DOCUMENT RESUME

ED 187 083 FL 010 927

AUTHOR Cohen, Andrew D.; And Others
TITLE Evaluating Evaluation. Bilingual Education Series: 6. Papers in Applied Linguistics.
INSTITUTION Center for Applied Linguistics, Washington, D.C.
PUB DATE Nov 79
NOTE 81p.
AVAILABLE FROM Center for Applied Linguistics, 3520 Prospect Street, N.W., Washington, DC 20007 ($5.75)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Bilingual Education; Criterion Referenced Tests; Educational Assessment; Educational Research; Elementary Secondary Education; Ethnography; Language Usage; Norm Referenced Tests; *Program Evaluation; *Reading Tests; Research Methodology; Speech Skills
IDENTIFIERS *Bilingual Programs; Natural Language

ABSTRACT

Results from three small-scale investigations of specific aspects of bilingual schooling as well as results of a more traditional large-scale evaluation of bilingual education programs in downstate Illinois are presented. The three small studies were designed to assess natural language in the classroom, compare criterion-referenced versus norm-referenced reading tests in bilingual classes, and provide an ethnographic description of bilingual language use patterns. An introduction to Evaluating Evaluation by Margaret Bruck and Andrew D. Cohen is included. The study articles are as follows: (1) "Assessing Bilingual Speaking Skills: In Search of Natural Language" by Andrew D. Cohen and Charles L. Roll; (2) "Assessing Bilingual Reading Skills: Criterion-Referenced versus Norm-Referenced Testing of Spanish Reading" by Flora V. Rodriguez-Brown and Andrew D. Cohen; (3) "Assessing Language Use in Bilingual Classrooms: An Ethnographic Analysis" by Margaret Bruck, Jeffrey Shultz, and Flora V. Rodriguez-Brown; and (4) "Assessing Bilingual programs: Methodological Considerations in Large-Scale Evaluations" by Andrew D. Cohen and Flora V. Rodriguez-Brown. Selected additions to the literature on bilingual education are included. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED187083

PAPERS IN APPLIED LINGUISTICS

Bilingual Education Series: 6

Evaluating Evaluation

by Andrew D. Cohen
Margaret Bruck
Flora V. Rodríguez-Brown

FL010927

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
The Center for Applied Linguistics
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

2

CENTER FOR APPLIED LINGUISTICS

ISBN: 87281-109-3
Library of Congress Catalog Card No.: 79-56013

November 1979

Copyright © 1979
By the Center for Applied Linguistics
1611 North Kent Street
Arlington, Virginia 22209

Printed in the U.S.A.

Contents

About Evaluating Evaluation: Introduction 1
Margaret Bruck & Andrew D. Cohen

Assessing Bilingual Speaking Skills: In Search of Natural Language 6
Andrew D. Cohen & Charles L. Roll

Assessing Bilingual Reading Skills: Criterion-Referenced versus Norm-Referenced Testing of Spanish Reading 22
Flora V. Rodríguez-Brown & Andrew D. Cohen

Assessing Language Use in Bilingual Classrooms: An Ethnographic Analysis 40
Margaret Bruck, Jeffrey Shultz, & Flora V. Rodríguez-Brown

Assessing Bilingual Programs: Methodological Considerations in Large-Scale Evaluations 57
Andrew D. Cohen & Flora V. Rodríguez-Brown

Selected Additions to the Literature on Bilingual Education 69

About Evaluating Evaluation: Introduction

Margaret Bruck
Andrew D. Cohen

The studies reported in this volume were planned and executed when issues for meaningful evaluation of bilingual schooling were coming into focus. Results from three small-scale investigations of specific aspects of bilingual schooling as well as a more traditional large-scale evaluation of bilingual education programs in downstate Illinois are presented. The three small studies were designed to assess "natural" language in the classroom, compare criterion- versus norm-referenced reading tests in bilingual classes, and provide an ethnographic description of bilingual language use patterns. These studies were not performed with the intention of evaluating specific bilingual education programs, but rather to furthering our theoretical understanding of issues involved in assessing oral language, reading, and language use skills. A second purpose was to document the actual procedures in performing the large-scale evaluation of one state's experience in bilingual education.

The volume as a whole should be of use to a wide-ranging audience. One target group includes researchers and evaluators, professors, graduate students, bilingual program evaluators, and psychometricians. A second includes administrators within school districts, coordinators of bilingual programs, and legislators who make the initial decisions about the implementation of research and evaluation and who then must decide whether or how to use the research for their purposes. A third group includes teachers in bilingual programs, teacher supervisors, teacher educators, and student teachers.

The main contribution of this book may be that it raises rather than settles issues relevant to research and evaluation of bilingual education programs. Hopefully, by virtue of the description of our experiences, readers will be sensitized to some of the problems as well as to some of the possibilities available to researchers in the field. In our planning and execution of the studies in downstate Illinois, four areas of basic concern became apparent, and careful consideration of these issues is crucial for *any* study of the phenomenon called *bilingual education*.

CLEAR DEFINITION OF INDEPENDENT VARIABLES

Few definitive statements about the success of bilingual education have emerged because of the elusive nature of this educational innovation. Many researchers have failed to carefully consider what bilingual education means in terms of the specific context they are studying. There is *no standard bilingual program,* and as pointed out in each chapter, the *treatment variable* is rarely constant between program classrooms or even within classrooms. Bilingual education encompasses many different approaches, both quantitative and qualitative. Approaches range from a class day where a large portion of the instruction is not in English to a situation where children attend a regular English class which they leave for 40 minutes of second language instruction.

Researchers must be cognizant of these variations. Furthermore they must be aware of the fact that while many programs may follow the same model (e.g., half-day pull-out with team teaching), it is rare that any two classes will follow this model in an identical fashion. Even within a classroom, the

curriculum or model may change (as is evident in Chapter Four), although the teacher or adminstrator might not be attuned to or report such changes. These examples point to the difficulty of considering pupils from a variety of bilingual classrooms as a homogeneous group which is often done in evaluations where children attending programs in various districts are grouped as *one bilingual group* and compared to a control group of children not attending a program. However, the differences among programs or classes within the bilingual group are often as great as the differences between bi- and nonbilingual programs. Thus, evaluators must carefully describe the parameters of the bilingual program under consideration, taking precaution that *all* the children are, in fact, following such a program.

The other kind of independent variable which must be carefully considered is the *background characteristics* of the children attending the programs. Often, rather than by program variables, results of bilingual evaluations can be largely explained by the student characteristics which were not properly controlled for in the original design. For example, in the large-scale evaluation reported in Chapter Five, we saw that the group with fewer years of bilingual education performed significantly different on many measures than the group with more bilingual experience. However, these differences probably are attributable in part to the fact that the first group was comprised of recent immigrants with less command of English than the second group, rather than to the lack of *bilingual* schooling per se. Some of the variables which must be carefully considered are the home language background (i.e. language skills and language use patterns), the subject's educational history, the socioeconomic level of the family, and culture-specific characteristics such as cognitive or personality styles, sex roles, etc.

In this context, the role of ethnographic data in any evaluation becomes an important one. Using these descriptive techniques, researchers will learn what is going on in the classroom so the sample of students can be properly selected and treatment variables accurately labeled. This type of data collection complements the more traditional use of other kinds of data which are also intended to identify the relevant characteristics of the sample and treatment variables.

In most evaluations, it will probably not be feasible to conduct an observational study at the level of detail presented in this volume. However, classroom interactions should be observed for several full school days throughout the school year. These observations of teachers, curriculum, and pupils can be conducted in fairly uncomplicated and meaningful ways. While observations have been conducted as part of many evaluational studies, unfortunately there is often no attempt to relate these to other testing procedures or testing results. In fact, in some cases the students who are observed are not the students who are tested. There must be a concerted effort to unite these two areas of evaluation--*classroom ethnography and summative, product assessment*--so that results can be interpreted in terms of *accurately perceived* independent variables.

CLEAR DEFINITION OF DEPENDENT VARIABLES

The main issue here relates to the question, "What *concepts* do we want to measure?" Too often, evaluators measure certain skills because of a test's availability not because of a theoretical or practical rationale for assessing the status of an ability or trait within the context of a bilingual education program. There should be a sound reason for including each variable and each skill or concept should be carefully defined before an instrument is selected. All too often because of ill-defined or inaccurate notions of the concepts being measured, we rely on the test constructors' notions of the concepts, but they may also be faulty. This approach was highlighted in Chapter Two where we tried to measure language dominance. The test did not fail to give an accurate measure of dominance; we encountered difficulties primarily because we had not carefully defined *dominance* before we began testing. Had we taken initial steps to define the concept, we might have perceived certain inadequacies in the instrument before giving it (and perhaps we would have added certain items to ensure an

adequate measure). Or, based on our notions of what constituted dominance, we may have been able to analyze the test data in such a way as to come up with an adequate measure of the concept. Likewise, evaluators should have detailed ideas about what they wish to measure when they talk about reading, language, self-concept, etc. Thus, evaluators are not seen primarily as technicians who fit together tests and labels of concepts, but as individuals who understand the process of education and have a theoretical and practical foundation in the areas of social, cognitive, and linguistic development and assessment.

SELECTION OF INSTRUMENTS

Once the concepts or dependent variables have been identified for study, then instruments must be selected to adequately measure these variables. Because of a lack of initial specification and description of the skills to be measured, because of a past dearth of bilingual assessment tools, and because of the present increase of new instruments often advertised with great zeal, evaluators are apt to accept an instrument without scrutinizing it carefully. Our study tested out two instruments, and several issues common to the evaluation of these tests were brought to light.

The first issue concerns the establishment of the *validity* of these tests. As mentioned, unless one has a good construct of the concept being measured, validation is often difficult to perform. Thus, in assessing the validity of a reading test, it makes sense to adopt the procedure of scanning the items to determine if, in fact, they demand behaviors or responses that are consistent with one's--or the test constructor's--notion of reading (i.e. establishing face validity). A second procedure is to establish *criterion validity* by correlating the test with another test commonly used in the field, as we did in the natural language study in Chapter Two. However, this procedure may result in a vicious circle since the reliability and validity of the second measure is, in many cases, suspect as well. A third way of validating an instrument is to determine whether the students' behavior which the evaluator is attempting to elicit through an item or task appears in a *naturalistic* setting. Frequently, children can be seen displaying the required response in a naturalistic situation, but not in the testing situation. At other times, it may be that the children do not produce the desired response because the behavior is irrelevant to their functioning. Often, a good measure of the validity of a test is *whether the measured behavior distinguishes more competent from less competent* children. So, in some instances, it may make sense to validate a test by classifying children a priori into high and low groups and then making every effort to observe the occurrence or non-occurrence of the specific behavior in a naturalistic setting.

The second issue in the area of instrumentation concerns the *role of classroom teachers* in the construction, administration, and scoring of tests. The BINL (Chapter Two) and the SOBER (Chapter Three) are examples of tests which try to involve the teacher in many phases of the assessment procedures--a feature which initially made these tests attractive to us. As evaluators we were aware of many teachers' feelings of isolation and disinterest in program evaluation. With good reason, teachers complain that their classroom schedules are disrupted by evaluators who spend long periods of time administering tests, questionnaires, or interviews to students and teachers, and then they rarely receive the results. As a result, many teachers have come to regard evaluation as another administrative procedure which has no relevance to their lives or to their students' lives. We hoped that we might overcome some of these negative attitudes through the use of the BINL and SOBER tests which involve the teachers in the evaluation process by having them design, administer, and analyze the tests. Another rationale for involving the teachers in the testing process was to add credibility to the test instruments. It was felt that were the tests or test items endorsed by an educational authority, that this would indicate the behavior under consideration and the means of testing it were educationally sound.

However, certain limitations to this approach became apparent. As was evidenced by the BINL, involving teachers in the actual data collection and analysis may not be a good evaluation strategy. For a variety of reasons, the teachers did not follow the prescribed procedures for collecting, transcribing, or scoring the natural language data. While these behaviors may be of interest for a separate study on teachers as program evaluators, they only interfered with the behavior we were trying to assess--the child's language skills.

Another role that the teachers played in our evaluation involved their examination and selection of items to be used on a reading test. While it is useful to have teachers' comments on the appropriateness of test items to ensure that the test measures relevant skills, the fact that teachers think a test is good does not automatically qualify it as a good instrument. In the SOBER study, teachers tended to like the test because items directly reflected skills taught by them. In the case of social studies or science, this judgment is probably of value. However, in the case of a reading test, a fit between what is taught and what is tested need not be so close. Children who have acquired a certain degree of reading skills should be able to generalize these skills and tackle material outside of the specific curriculum in the classroom. If they cannot, then they probably cannot read very well. The objectives that teachers stress in their reading program are, after all, just a means to attain a more general goal.

It may also be that the assessment needs of teachers and of program evaluators are different, and that for evaluators, the use of norm-referenced tests is advisable while the use of criterion-referenced tests is advisable for teachers. For example, the testing of items specifically related to the syllabus *does* give teachers feedback on each pupil's attainment of material taught in the teacher's program. Yet, it may be the case that the current wave of interest in testing integrative language skills will once again put the emphasis on more general, standardized tests of reading and that the pendulum will swing away from criterion-referenced testing. Clearly more work is needed in this area to see where criterion-referenced tests can benefit teachers, evaluators, and researchers.

In general, we feel that we must *reassess our ideas of what we mean by teacher involvement* in the evaluation procedure. Our experience indicates that teachers should not be given the roles of research assistants. Perhaps, their participation would be more useful in the planning stages of the evaluation such as defining the concepts to be measured, describing the specific characteristics or parameters of their bilingual classes and students, as well as helping to interpret the results once they have been analyzed. Devising ways for teachers to have input into the evaluation, as well as guaranteeing procedures to make them aware of the results, should be a basic component of any program evaluation.

SELECTING AN APPROPRIATE SCOPE FOR RESEARCH AND EVALUATION

Perhaps the most interesting and informative studies in this project were the three pilot studies. We attribute this to the manageability of the studies in terms of the number of subjects and skills assessed. Because the sample sizes were small, we could easily control for background characteristics such as language background, age, grade, and bilingual program model. Furthermore, we could carefully monitor testing procedures to ensure appropriate administration and to detect specific difficulties that arose during the testing periods. Because the range of skills was limited to one area in each study, we did not feel swamped by the data and could spend time intensively analyzing those available to us.

Unfortunately, too many evaluations have tried to attack all areas in one massive attempt. This has involved the expenditure of large sums of money, the testing of hundreds or thousands of children, and the use of many different tests and testing procedures. The results of such evaluations are usually uninterpretable because of poor testing procedures, poor control of independent variables, inadequate sample selection techniques, missing data, and inappropriate statistical analyses.

Our experience with the Illinois data as reported in Chapter Five is characteristic of those school districts or evaluations that begin large-scale evaluation with no real blueprint for action.

We urge evaluators to *start small,* to select one or two classes for initial study, to become familiar with the curriculum, the teachers, and the students, and to select a few critical areas for testing. Analyze the data as soon as possible, and produce a final report within a few months of the end of data collection. At this point, review the data with the appropriate education personnel, and with their help, *plan the next evaluation* which, with each passing year, may expand to the testing of different skill areas using different types of students in different programs.

This plan of action is, in fact, one that has been followed by two small-scale evaluation projects which, over the past 10 to 15 years, have had an enormous impact on our knowledge of the consequences of bilingual education. These are the St. Lambert French Immersion project in Montreal and the Culver City Spanish Immersion project in southern California (projects with which two of the current authors, Bruck and Cohen, have been involved). With both of these projects, one or two classes with appropriate controls were tested on a yearly basis. Research questions for each evaluation were either determined by educators' and parents' concerns, based upon clarification of previous results, or stemmed from theoretical interests of specific researchers. Actually, a research interest of a particular researcher has often added a fresh theme to many evaluations and has led to nontraditional evaluation which reflected some current concern in a specific field. As a result, the horizons of the evaluation have been broadened. With each passing year, new and additional groups have been tested so that presently there is a fairly comprehensive research base on which to reach some definitive conclusions about the effects of bilingual education in these specific settings.

The typical American experience in evaluating bilingual programs has consisted of much testing with few tangible insights gained as to the effects of bilingual education. Perhaps smaller-scale, more methodologically tight evaluations would be met with greater credibility and would play a more conspicuous role in the drafting of bilingual education policy. It would be particularly rewarding to evaluators if decision makers would use evaluation to determine program design, rather than to draw from evaluation on a *post facto* basis, if at all, and possibly to justify a course of action already chosen.

After 10 years of federally funded programs, we still know surprisingly little about the academic, sociocultural, or linguistic consequences of American bilingual education. We hope that this volume is a fresh start at honestly looking at some of the problems that must first be resolved before larger studies can be undertaken.

Assessing Bilingual Speaking Skills: In Search of Natural Language

Andrew D. Cohen
Charles L. Roll

Since the development and implementation of bilingual education more than a decade ago, bilingual language assessment has remained an acute concern of individuals involved in all phases of bilingual programs. While tests have been developed (see Stolz and Bruck, 1976, for a review), it is questionable whether any one test adequately measures the three critical areas of bilingual language assessment: (a) language proficiency--pupils' skills in each of their languages, both productive (speaking and writing) and receptive (listening and reading); (b) language dominance--pupils' stronger language for one or all of the skills; and (c) language change--pupils' increase or decrease of skills in one or both languages over time (e.g., improvement in speaking one language, maintenance or loss in speaking the other). In reality, we are not at all sure as to the best way to tap each of these areas; there may not be a *best* way, only possibly, a *better* way.

A good testing instrument for overall language assessment in bilingual programs should be one that gives a reliable and valid measure of language proficiency and of language change over time. In order to determine language dominance among bilingual children, a good instrument should also have to account for differences across the two languages. For example, a test of specific grammatical structures aimed at determining dominance should account for differences in the significance, frequency, and difficulty of such structures across languages. We explored the problems of developing such a test by piloting an assessment instrument which claims to measure bilingual language proficiency, dominance, and change--the *Basic Inventory of Natural Language* (BINL).

We were initially drawn to the BINL because of several features which made it look like a good test. A principal feature of it, and a *major* reason why the instrument was selected for this study, is its explicit procedures for tapping "natural" or freely given speech. In this test, language is not elicited through structured or semistructured questions, but rather through activities which are considered natural within the context of a small group in a classroom.

Another feature is that it is a "complete criterion-referenced system designed to assess the language dominance, proficiency, and growth of students in school language arts and reading programs (Herbert, 1977)." In other words, the instrument is to be used both as a norm- and as a criterion-referenced test. As a norm-referenced test, student results are to be compared to those of other students within the same classroom, across classrooms within the same school, across schools and districts, etc. As a criterion-referenced test, the classroom teachers are to evaluate individual students' performance in relation to fixed criteria--in this case, the frequency with which students use particular grammatical structures. After individual "student profiles" are completed, the teachers and their aides are directed to prescriptive activities found in the BINL manual (Herbert, 1977). These activities are meant to give students the opportunity to practice and develop their dominant and second languages in view of individual profiles.

The purpose of our study was to give the BINL as outlined in the manual, and to determine how good a measure of language it was. Although the focus is on one test, we believe the present approach to evaluating the strengths and weaknesses of this test is generalizable to other tests:

- *In what ways does the test determine oral language proficiency and how well does it do this?*
- *How well can the test be used to determine language dominance?*
- *Is the test sensitive to pupils' language change?*
- *Can the test determine comparative development in two languages?*
- *How well does the test differentiate the more proficient speakers from the less proficient ones?*
- *To what extent do results from the test correlate with results from another measure of oral language proficiency?*
- *What relationship does the test scoring system have to an error analysis approach to the same data?*
- *What are the characteristics of the "natural" language assessed by the test?*
- *How well suited is the test for teacher administration and scoring?*

THE BINL PILOT STUDY

Subjects

The subjects were elementary schoolchildren participating in tutorial pull-out bilingual programs in three multigraded classrooms in a school district in downstate Illinois. The students spent from one to five hours a day in this classroom, with a Spanish-English bilingual teacher. Twenty children were selected for the pilot study, five each from grades two through five. We omitted students who had little or no knowledge of one language so as to obtain a more valid assessment of the BINL as a bilingual instrument. Three were ultimately removed from the fourth grade sample because their pretest samples were lacking the necessary corpus of 10 sentences. Thus, 17 subjects comprised the final sample which was purposely kept small for detailed analyses.

Measures

The BINL kit contains materials to elicit language from children, including static pictures ("Story Starter Posters"), sequence pictures ("Story Sequence Pictures"), and groups of tiles ("Talk Tiles"). Students are to be well acquainted with these elicitation materials, with the tape recording device,* and with the whole procedure of group language activity before samples of language are used for assessment purposes. (At least 20 minutes of familiarization a day for one to five days, is recommended before the stories are used as data.)

In obtaining the data, students are trained to talk about individual pictures or sequences of pictures to other children in groups of six or less. Adults are not supposed to use questions to prompt children to speak during the gathering of the language sample, except for those questions "that occur as a part of the natural language interchange between students...[but] whenever possible it is best to eliminate responses to questions and only use narrative as language samples (Herbert, 1977)." This procedure is intended to replace an adult examiner questioning a child on a one-to-one basis, thus diminishing any inhibitions children might have about speaking in either language. With regard to inhibition in formal testing situations, Labov (1969) demonstrated the dramatic extent to which a more informal testing situation can change the nature of the data collected.

**Before data collection, student groups are instructed in the use of sound page machines--devices which tape record four minutes of language onto special "pages" that can be easily manipulated, even written on and folded.*

A sample of language is to consist of 10 sentences (partial and/or complete) (Herbert, 1977). Teachers can either score the samples on their own or can tally the number of words in each sample and then submit the samples to machine scoring (machine scoring was not available at the time of this study). If the samples are machine scored, teachers still need to examine the individual language samples of the students in order to recommend prescriptive activities for students relative to their language development needs (Herbert, 1977).

The BINL system for the assessment of an individual's language skills provides a series of scores intended to represent the profile of each student's oral skills in Spanish and English. (Work on other languages such as Portuguese is underway.) One of the scores of the profile is that of *Fluency*--the total number of words used by the student in a sample of sentences. This score is considered to be a general indicator of ability to use the "vocabulary, structures, and forms of a language." A second score is that of *Average Sentence Length*--the total number of words divided by the total number of sentences (generally 10). A third score is that of *Level of Complexity*--a count of the number of partial sentences, independent clauses, dependent clauses, modifiers, and phrases found in the set of consecutive sentences within the corpus. These elements are weighted according to a predetermined notion of the extent of proficiency or development that the elements reflect. A partial sentence (e.g., "trying real hard") is worth one point, an independent clause (e.g., "Jack lives on my street") is worth 10 points, a modifier (e.g., "the," "pretty," "now") is 10 points, a phrase (prepositional, "on the hill"; adverbial, "as quick as a rabbit"; participial, "holding the book"; infinitive phrase, "to play") is 20 points, and a dependent clause (e.g., "because he likes animals, he owns three dogs") is 30 points. These points are summed within and across sentences to produce the *index of language ability*. The average level of complexity is this sum divided by the total number of sentences (in this case, 10). (The following pages contain samples of both BINL score sheets with sentences from both English and Spanish.) Language dominance is determined by subtracting the difference between the BINL score in the two languages. For the purposes of this scoring system, it does not matter whether the sentences are grammatically correct or incorrect according to adult norms for standard language. The difference between total pre- and post-test scores (e.g., average level of complexity) is intended to serve as a measure of language "development"--what we refer to as language "change," in that language skill can be improved, maintained, or lost over time. (At the time of the study, reliability coefficients were not available, either for English or Spanish samples.)

In addition to the scores obtained from following the BINL manual, we also calculated a set of error analysis scores. These served as a standard to evaluate the validity of the BINL measures. Error analysis categories for the speech samples were: grammaticality of the sentence; the concordance between determiner and noun, between adjective and noun, between subject and verb, between subject and verb phrase; subject pronouns, direct and indirect object pronouns; prepositions; relative clauses; and subordinate clauses. A student's score for each category consisted of the ratio of deviant forms to total forms occurring in that category. All categories were then summed to produce a total score. This analysis was performed on the same 10 sentences selected for the BINL scoring as well as on the total corpus of sentences.

The other oral language proficiency measure used in the study was the MAT-SEA-CAL, developed for levels K through four by Matluck and Mace-Matluck, the Seattle Public Schools, and the Center for Applied Linguistics. Since this test claims to measure language proficiency and language dominance, and to provide a diagnostic analysis or profile of the student, it was used to validate the BINL. Language dominance is determined both by comparing proficiency in each language and by examining the diagnosis for each language. The test has appeared in Spanish, English, Cantonese, Mandarin, Tagalog, and Ilokano. The reported Kuder-Richardson 20 reliability coefficient for the English version is .94, and for the Spanish, .91. The authors report that no attempt was made to validate the test by

INDIVIDUAL ORAL SCORES SHEET (IOS)

BASIC INVENTORY OF NATURAL LANGUAGE (BINL)

Student's Name _____

Grade 4 Teacher _____

Language English

☐ PRE-TEST _____ DATE

☒ POST-TEST _____ DATE

I. Simplified Scoring

	A Complete Sentence	B Partial Sentence	C Fluency	Samples	Grammar Analysis Columns: Partial Sentence (Points 1)	Independent Clauses (10)	Modifiers (10)	Phrases (20)	Dependent Clauses (30)	Sub-Total
	1		13	One day there was a goat watering his flowers, so they'll grows better.		10	40	20	30	100
	2		4	A lion came along.		10	10	0	0	20
	3		9	He was so thirsty, that he couldn't find water.		10	10	0	30	50
	4		13	it took the hose away from the goat and drank all the water.		20	40	40	0	100
	5		12	Poor little goat, he doesn't had any more water for the flowers.		10	40	20	0	70
	6		3	The flowers died.		10	10	0	0	20
	7		4	He was thirsty too.		10	10	0	0	20
	8		6	A mice came along eating cheese.		10	10	20	0	40
	9		11	The mouse was thirsty too, because he was eating too much.		10	40	20	30	100
	10		8	The goat got more water in a vase.		10	30	20	0	60
	10		83	Totals						580
	A	B	C							

Total Index of Language Ability 580

II. Syntactic Analysis (Hand Scoring)

		NUMBER OF:
83	Total number of words: (Total of Column C)	PARTIAL SENTENCE
10	Total number of samples: (Enter sum of ✓ marks in Columns A & B)	INDEPENDENT CLAUSES
8.3	Average Sentence Length: (Divide total number of words by total number of sentences)	MODIFIERS
58	Average Level of Complexity: Divide Index of Language Ability (above right) by Total number of sentences	PHRASES
		DEPENDENT CLAUSES

INDIVIDUAL ORAL SCORES SHEET (IOS)

BASIC INVENTORY OF NATURAL LANGUAGE (BINL)

Student's Name

Grade 2 Teacher —

Language Spanish

☐ PRE-TEST DATE

☒ POST-TEST DATE

I. Simplified Scoring

Grammar Analysis Columns

	A Complete Sentence	B Partial Sentence	C Fluency	Samples	Points	1 Partial Sentence	10 Independent Clauses	10 Modifiers	20 Phrases	30 Dependent Clauses	Sub-Total
	1		7	Un día, un chivo quería tener flores.			10	20	0	0	30
	2		3	Les puso agua.			10	0	0	0	10
	3		1	Crecieron.			10	0	0	0	10
	4		8	Cuando se fue a (a) dormir, era la noche.			10	10	20	30	70
	5		9	Cuando vino la mañana, el conejo rompió las flores.			10	30	0	30	70
	6		3	Se los comió.			10	0	0	0	10
	7		5	Se fue pa' la casa.			10	10	20	0	40
	8		9	El chivo lloró todo el día por sus flores.			10	20	40	0	80
	9		6	Fue a la casa del conejo.			10	10	40	0	60
	10		7	Dijo: "Por qué te comiste las flores?"			10	10	0	30	50
	10		58	Totals							430
	A	B	C								

Total Index of Language Ability 430

II. Syntactic Analysis (Hand Scoring)

		NUMBER OF:
58	Total number of words: (Total of Column C)	PARTIAL SENTENCE
10	Total number of samples: (Enter sum of ✓ marks in Columns A & B)	INDEPENDENT CLAUSES
5.8	Average Sentence Length: (Divide total number of words by total number of sentences)	MODIFIERS
43	Average Level of Complexity: Divide Index of Language Ability (above right) by Total number of sentences	PHRASES
		DEPENDENT CLAUSES

correlating it with other oral proficiency or dominance measures (Matluck and Mace-Matluck, 1976).

The MAT-SEA-CAL is administered to children individually by an adult tester and is recognized to be a measure of formal language in the educational domain. The authors have set a criterion for determining dominance, namely a 10 percent or greater difference in total score (maximum 100 percent) between versions in the two languages. If the scores are within 10 points of each other and both are over 70 percent then the pupil is to be referred to as "high balanced."

The subtests used in the study were: *Listening Comprehension* (27 items)--for the first 17 items, the examiner says a sentence and the child points to the picture from among four which best "gives the meaning." For the last 10 items, the examiner gives a command to which the child is to respond (e.g., "stand up"). *Structured Response* (28 items)--the respondent is shown a picture and asked a question about it. The question is designed to elicit specific grammatical structures. The answers are scored for grammatical correctness, not for correct information.

Although both are oral language proficiency tests, the BINL and the MAT-SEA-CAL are, in many ways, divergent means of testing the same construct. First, the BINL is intended to be a "direct" (natural) approach to assessment in that it supposedly samples directly from the behavior being evaluated (in this case, children telling stories to their peers). The MAT-SEA-CAL is considered "indirect" (artificial, structured) in that the assessment is contrived and thus taps something different from the ability under consideration (in this case, elicitation of specific grammatical structures is equated with language proficiency, cf. Stolz and Bruck, 1976). Second, although the BINL and the MAT-SEA-CAL are both discrete point tests, the scope of the scoring differs. The BINL uses broad syntactic categories as the discrete items (a count of independent and dependent clauses, modifiers, phrases), whereas the MAT-SEA-CAL uses very specific items and grammatical relationships and structures; the BINL ignores grammaticality in responses, whereas the MAT-SEA-CAL's Structured Response Test (the oral production test) scores only for grammaticality. Third, the BINL does not have a receptive or comprehension section like the MAT-SEA-CAL.

Data Analysis

The BINL data (collected over a five month period) were to be transcribed and scored by the teachers; all the sound page samples were then re-recorded onto cassette tapes and transcribed by a trained linguist; these were then scored according to the BINL system. This procedure was intended to provide a check on the teachers' transcriptions and scoring methods.

In working with the BINL data, it became apparent that a more detailed analysis of language development was needed, than simply the difference between total pre- and post-test scores. Hence, the raw frequencies within each grammatical category (e.g., independent clause, phrase, etc.) were summed individually both at pre- and post-test assessment and then compared. In addition, the raw frequencies rather than the weighted points were summed in order to give a clearer picture of how often forms actually occurred. In order to compare group performance in Spanish and English for the BINL pre- and post-test assessments, means and standard deviations across the 17 students were computed.

Pearson correlation coefficients were computed to determine the relationship between the BINL and MAT-SEA-CAL results using the SPSS program (Nie et al, 1975). Intercorrelations of the two BINL scores, average length of complexity and average length of sentence, were so high that it was deemed adequate to use just the average length of complexity score to represent the BINL in the correlations.

The Pearson correlation statistic was employed to relate the BINL score to the total error analysis score on the same spoken corpus. The error analysis score used for this study was that for the 10 sentences within the BINL sample. A separate error analysis was calculated on the entire corpus of data for each child

for the pre- and post-test in both languages. However, we found that the results for the entire corpus were similar enough in terms of proportion and type of errors to the results for the 10 sentences alone to warrant the use of the error analysis of the BINL portion alone.

FINDINGS

●*In what ways did the BINL determine oral language proficiency and how well did it do this?*

The BINL manual defines proficiency as "the ability to use the various components of language such as vocabulary, structure, and morphology to express one's thoughts (Herbert, 1977)." In actuality, the instrument only assesses the existence and frequency of certain structures and structure words (e.g., modifiers) in children's speech. The BINL does not assess richness of vocabulary (i.e type-token analysis) nor does it assess morphology (e.g., correct word formation, past tense markers, pluralization). Whereas the BINL system places emphasis on the freedom that the students have to choose their own stimuli for stories and the unstructured way in which they are to tell their stories, the scoring system, in fact, focuses only on certain discrete categories. Consequently, students telling a story about "a boy and all the children" obtain more points for "modifiers" than students telling a story about themselves ("I"). Every time the student says "all the children," 20 points are awarded, 10 for each modifier ("all" and "the"), whereas "I" yields no points for modifiers. In the Illinois study, one third grader seemed to *regress* in modification from pre- to post-test (170 points on the pre- and only 90 points on the post-test), but the subject of the story shifted from "all the children" to "I." Thus, the apparent loss was actually an artifact of the scoring system.

Furthermore, no rationale is given for the weighting of grammatical categories within the scoring system. The weights *seem* to reflect levels of grammatical development, such that phrases are considered at a slightly higher level of development than modifiers, and so receive greater weighting. There are several problems with such a weighting system. First, grammatical complexity does not always parallel language development in that grammatically complex, but semantically salient forms (e.g., dependent clauses introduced by "because," 30 points) may develop before grammatically less complex, but semantically less salient forms (e.g., participial clauses introduced by a gerund, 20 points) (see Slobin, 1973). Second, disparately different structures within the same grammatical category are weighted the same. To illustrate the problem of differences within categories, let us take the BINL category "dependent clauses," worth 30 points. Whereas relative clauses introduced by *that* and *who* and adverbial clauses introduced by *when* seemed to predominate in the Illinois English language data, there was only one *if* clause. However, all were given the same weight. Might not the appearance of an *if* clause reflect greater semantic sophistication, namely the notion of the conditional, and thus be given a greater weight? Furthermore, and perhaps of greater consequence in scoring, direct quotes were scored as dependent clauses, and thus each received 30 points. Yet, direct quotes actually seemed to show less syntactic skill than indirect or reported speech, which involved several transformations. A pupil's pretest score for dependent clauses was sometimes higher than the post-test score only because in the pretest three or four direct quotes were used, while in post-test one or two subordinate clauses were used. Thus, without looking at the actual data, the scoring system would mislead the investigator into thinking that the child had regressed in his grasp of dependent clauses.

The BINL is, then, a measure of oral language proficiency, but only with respect to average sentence length and frequency of specified grammatical categories. These are areas of analysis that other oral proficiency instruments have also focused attention on, so there is some agreement among test developers that such indices measure oral proficiency. However, it could be argued that shorter sentences may show greater fluency than longer, more incoherent ones (particularly

where grammaticality of sentence is ignored) (see Swain and Bruck, 1976).

These analyses highlight the importance of the researcher's scrutinizing of the measures used in many language tests which purport to assess overall fluency. This not only means a careful reading of manuals, but also careful examination of actual data collected.

●*How well did the BINL determine language dominance?*

Because the BINL manual did not provide any formal criteria for judging dominance, nor were any readily apparent from the data analysis, we had to establish our own criteria. A detailed analysis of the data across languages brought to light several issues. First, assuming an equivalence in the scoring systems between English and Spanish, what point differential would be necessary to establish dominance? (This is actually *not* a small assumption to make. Whereas comparability of scores across languages is of critical concern, the BINL manual does not discuss parity of scoring forms at all.) What measure should be used? As mentioned, the BINL manual did not provide guidelines here. Perhaps because of certain basic similarities between Spanish and English, BINL subscores by grammatical category seem reasonably comparable. However, because Spanish allows certain optional deletions of word classes, such as subject pronouns, it might be expected that average sentence length in Spanish would be shorter. To test for comparative length, we located students whose Spanish and English BINL average level of complexity scores (from either pre- or post-testing) were within five points of each other (N=15), presumably indicating no dominance of one language over the other. We then calculated the average difference in sentence length. English sentences were found to be more than a half a word longer on the average (.63). Actually, of the 10 differences that favored English, one corpus had an average of 3.1 more words in English than in Spanish. Of the five that favored Spanish, the greatest difference was an average of one word more. Over large numbers of students, this difference could be quite significant. Because this finding suggests a lack of parity between languages on this one index--sentence length--one should be cautious in using such an instrument for between-language comparisons.

The next attempt to determine a measure of dominance involved calculating the difference between the 17 pairs of pretest level of complexity scores. In only six cases was the difference in average level of complexity scores 10 points or greater. Three of these differences favored Spanish and three favored English. We checked these students' MAT-SEA-CAL Spanish and English scores as a measure of convergent validation and found that they were generally in agreement with these assumed dominance patterns. At this point we might have been able to make some statements about apparent dominance. But then, we also looked at teachers' ratings of the language dominance of their students (Spanish, English, or balanced), and found that for three of the six students, the teachers' ratings were not consistent with BINL and MAT-SEA-CAL findings. In two cases, the teacher rated students as English dominant when BINL and MAT-SEA-CAL scores favored Spanish, and in one case a teacher rated a student as Spanish dominant when the student scored higher on the English versions of the tests. The other three teacher ratings matched the test scores.

What explanation is there for this lack of consistency? Perhaps teachers' notions about student language dominance are more impressionistic than those of tests based on carefully selected linguistic factors. Or, perhaps teachers put more emphasis on different linguistic elements than the tests do, i.e. the element of pronunciation. Then again, it may be that the tests are underrating students who can communicate quite well in one language, but lack some of the structural features of that language. There is even the possibility that tests such as the BINL and the MAT-SEA-CAL do not tap the *more* natural language environments that the teachers have access to every day.

These findings suggest that assigning language *dominance* is not an easy matter, and that perhaps more of the onus should be on test constructors to provide rigorous, valid criteria for identifying the dominant language.

●*Was the BINL sensitive to pupils' language change?*

The BINL only provided assessment of change in certain broad grammatical areas. Thus, while the instrument did deal with determiners and with different types of clauses, it did not deal with the development of verb tenses, negative questions, and negation, all of which are areas of specific interest to second language acquisition researchers. Within the BINL categories, it was possible to observe changes from pre- to post-tests. Needless to say, it is important to allow enough time between two assessment periods for there to be real development. In this study, there were approximately five months between pretest (November-December) and post-test (May). Such a period was probably adequate. Actually, these "middle" months of school may have been the best time to discern gains, since the first two months of schooling are often spent in recovering losses from summer recess (see Cohen, 1974) and the last month or so of school reflects a winding down before the summer.

As in the case of assessing dominance with the BINL, the point score can only be used as a rough indicator of change; the user must also look closely at the behavior that the scores reflect. For example, there were seven instances where a change of about 10 points in the BINL average level of complexity score from pre- to post-testing did not appear to reflect a genuine change in language proficiency, but rather reflected an artifact in the scoring procedures as discussed. Where careful analysis of the data suggested that genuine change was taking place the gain or loss averaged 21 points.

In doing a category-by-category analysis of the total scores, we found that changes were not necessarily caused by an increase in all categories. In one case, a second grader had a dramatic increase in English, primarily in the use of modifiers, from 10 in the pretest sample to 32 in the post-test. Another student, a third grader, doubled the number of phrases in Spanish from pre- to post-testing (8 to 19), while gains in other categories were more modest. Hence, the validity of judging language change, when the claim is based on a differential pattern of occurence of *one* grammatical structure is questionable.

Furthermore, an instrument with structured categories of analysis will only produce direct assessment of change within those categories. An ideal measure of language development would provide guidelines for evaluating change based on research on the developmental sequence of structures in the acquisition process among specific learner groups.

●*Did the BINL determine comparative development in two languages?*

Comparisons of bilingual development were possible using the BINL and were meaningful to the extent that such cross language comparisons could be considered valid. With respect to *bilingual* development in level of complexity, two students made gains in both Spanish and English, four made gains in Spanish without any change in English, four made gains in English without any change in Spanish, and six did not change in either language. One student lost ground in Spanish and stayed the same in English. Table 1 on page 16 presents these data graphically. The Table suggests that among these second and fifth graders, English and Spanish were being maintained and that gains were being recorded in one language more so than in both, with such gains as likely to be in favor of Spanish as English. Of course, the sample is so small that the findings must be little more than suggestive of future evaluation procedures.

Another way of looking at the data is in terms of group means for level of complexity by language over time. Here again with these data, the results are only suggestive (see Table 2, page 16). It can be seen that average gain in English was greater (eight points) than average gain in Spanish (3.6 points). Furthermore, the standard deviation in Spanish was generally higher than in English, suggesting greater dispersion from the mean among Spanish scores than among English scores.

Thus, it appears that the BINL might provide insights into the relative development of two languages by *individual students*. But any such comparisons should be based on close order analysis that considers the extent of parity

between the scores across languages. Given the small scale nature of this study, it is not really possible to adequately evaluate the use of the BINL in assessing comparative language change on a *group* basis.

●*How well did the BINL differentiate the more proficient speakers from the less proficient ones?*

Teachers using language complexity as a measure of proficiency could obtain a rough measure of the range of proficiency within their classes from the BINL. For example, the third grade Spanish pretest scores showed a range of 35 points (33-68), while the English post-test range was only eight points (52-60) (N=5). These results would suggest to the teacher that Spanish proficiency was more varied than was English proficiency within this particular classroom. Several measures of Spanish language complexity taken every few weeks could yield a reliable measure of a pupil's performance relative to other students in the same class. One test score alone, however, may be less reliable, particularly if it is intended for use in tracking or grouping students, in that the error variance or deviation from the student's "true" score may be great. Error variance for one BINL score could result from the fact that the instrument does not control for the stimulus (i.e. the pupil may have more vocabulary or be more motivated to talk about topic *X* as opposed to topic *Y*) or the instrument is subject to the problem of scorer reliability (i.e. the scorer includes too many segments in one sentence--beyond the natural or stipulated sentence boundary, thus inflating the student's score).

This study did not really provide enough data to suggest whether the BINL is sensitive enough to pick up differences *across* grades. It was true, however, that longer sentences (i.e. 10 or more words) only appeared at the fifth grade level, and not below. There was *no* progressive trend in BINL complexity of language scores from grades two to four. In fact, the highest Spanish post-test score from grades two to four, 74, was obtained by a second grader. Obviously, there are criteria other than grade, related to sentence length, namely, chronological age and mental age or intelligence.

Thus, the BINL does provide a rough measure of comparative proficiency across students although the availability of norms (not available at the time of this study) would add credibility to any cross-student comparisons. Nonetheless, teachers may want to consider using an instrument such as the BINL in order to learn how the pupils in their classrooms compare to one another in one language or both. Administrators and researchers may wish to determine oral language performance of students at different grade levels in a given school or across schools. Such a question is of particular importance to administrators of schools with large immigrant populations, particularly where new immigrants are arriving all the time and are at all grade levels.

●*To what extent did results from the BINL correlate with results from another measure of oral language proficiency?*

In general, the MAT-SEA-CAL subtests were all relatively easy for the students. The mean English scores were also generally higher than the Spanish scores, roughly eight points higher on listening and seven points higher on structured response (see Table 3 on the following page).

The MAT-SEA-CAL Spanish listening comprehension subtest had a nonsignificant correlation of .36 with the BINL Spanish pretest score (i.e. level of complexity) and a .03 correlation with the BINL Spanish post-test; the MAT-SEA-CAL English listening comprehension subtest had a significant ($p < .05$) *negative* correlation (-.41) with the BINL English pretest, and a -.33 correlation with the BINL English post-test (Table 4 following). The one significant correlation appears contradictory since we might expect more proficient listeners to be more proficient speakers as well. All the same, the sample size (17) and the level of significance (.05) suggest that this finding may be a statistical artifact rather than a genuine trend.

TABLE 1

BINL Change in Bilingual Skills

ENGLISH \ SPANISH	Gain	No Change	Loss	
Gain	2	4	0	6
No Change	4	6	1	11
Loss	0	0	0	0
	6	10	1	17 (Total N)

TABLE 2

BINL Average Level of Complexity (N=17)

	ENGLISH		*SPANISH*	
	Pretest	Post-Test	Pretest	Post-Test
Mean	54.6	62.6	54.9	58.5
S.D.	12.6	12.7	15.1	14.1

TABLE 3

MAT-SEA-CAL Results (N=17)

		ENGLISH	*SPANISH*
Listening Comprehension	Mean	95.9	87.7
	S.D.	4.8	6.0
Structured Response	Mean	93.5	86.3
	S.D.	5.7	6.9

TABLE 4

BINL and MAT-SEA-CAL Correlations (N=17)

MAT-SEA-CAL \ BINL	*SPANISH* Pretest	*SPANISH* Post-test	*ENGLISH* Pretest	*ENGLISH* Post-test
SPANISH Listening Comprehension	.36	.03		
SPANISH Structured Response	.40	.29		
ENGLISH Listening Comprehension			-.41*	-.33
ENGLISH Structured Response			.18	.25

*$p < .05$

Correlations between the MAT-SEA-CAL structured response subtest and the BINL average level of complexity were all positive, but not significant. The MAT-SEA-CAL Spanish subtest had a .40 correlation with the BINL Spanish pretest and a .29 correlation with the BINL Spanish post-test. The MAT-SEA-CAL English subtest had a .18 correlation with the BINL English pretest and a .15 correlation with the BINL English post-test (Table 4).

The findings suggest that the MAT-SEA-CAL and the BINL average level of complexity index are testing different aspects of oral language fluency, with the MAT-SEA-CAL listening comprehension subtest diverging more from the BINL than the grammatical structure subtest. Another explanation is that neither test is really getting at "language." Ideally, a validated test that could be used as a criterion measure to assess the validity of each of these instruments should exist. The lack of such an instrument in this case prevents any clear-cut statements about criterion validity.

•*What relationship did the BINL scoring system have to an error analysis approach to the same BINL data?*

Generally there was a significant correlation between the BINL and error analysis results. The two scoring systems thus appear to be measuring related behaviors. The BINL English pretest average level of complexity scores and the error analysis total scores correlated highly significantly (r=.75, p<.001). The BINL Spanish pretest complexity scores and error analysis scores also correlated at a relatively high level of significance (r=.61, p<.01). The Spanish post-test correlation was statistically significant, but not as high (r=.47, p<.05), and the English correlation (r=.15) was nonsignificant. The findings suggest that in pretesting, students with more complex language also made fewer mistakes. This tendency, although less pronounced, was also found with the Spanish post-test.

Error analysis was conducted on BINL data primarily to determine how these findings would relate to the findings produced by the BINL scoring system as described in the manual. The results showed that the two kinds of analysis complemented each other in that the error analysis alone might penalize students for incorrectly producing a subordinate clause, while the BINL would award 30 points for the clause, regardless of how well formed it was grammatically. Just as the BINL could be used judiciously to assess change in average level of complexity and in average sentence length, so the error analysis approach can be used judiciously to determine changes in grammatical control--with respect to the norm for native speakers. In the current data for example, there was one instance of enhanced mastery of Spanish grammar over time. In pretesting, the student's numerous errors involved basic agreement--primarily subject-verb, noun-determiner, noun-adjective, etc. In post-testing, there were no problems in these areas. Problems concerned other, less basic forms, such as a subject-indirect object pronoun relationship and a subordinate verb phrase. In the case of another student, errors in Spanish increased from pre- to post-test. In pretesting, there was only one prepositional error, but in post-testing there were a number of agreement problems--subject-verb, noun-determiner, noun-adjective, etc. The Spanish level of complexity score for this student also dropped from pre- to post-testing, from 77 to 46.

The pretest Spanish and *both* the Spanish and English post-test error analysis scores *also* correlated significantly with the scores on the MAT-SEA-CAL structured response subtest: English error analysis pretest with structured response, r=.17, English error analysis post-test with structured response, r=.50, (p<.05); Spanish error analysis pretest with structured response, r=.63 (p <.01), Spanish error analysis post-test with structured response, r=.60 (p <.01). Error analysis scores did not correlate significantly with listening comprehension. These findings would suggest that both the structured response and the error analysis are assessing similar things. This is not surprising since the structured response subtest is scored according to whether the sentence is completed *grammatically*. The listening comprehension subtest, on the other hand, was not primarily designed to tap grammaticality.

Thus, it would appear that linguistic investigation using error analysis techniques to assess acquisition of normative language, can provide a useful complement to simpler procedures which do not assess deviations from the norm.

●*What were the characteristics of the "natural" language assessed by the BINL?*

An attempt to answer this question simply produced more questions. First, how easy is it to remove the adult from the testing session? In this study, it appears that the teachers made their presence felt during each data collection session. The students were clearly aware that the adult was there, and in fact, the adults intervened to elicit language from the children in a number of instances. Second, how "natural" is a storytelling situation? There may, for example, be a basic difference between the language of storytelling and, say, *casual* classroom conversation with peers, with the latter being more "natural." Not only do topics discussed in casual conversation differ from those treated in stories, but also the very nature of the discourse is different.

A sample of data collected as part of the language use study reported in Chapter Four consists of a transcription of a videotaped portion of dialogue (mostly in Spanish) between a Puerto Rican bilingual first grader, Rosa wearing a wireless microphone, and a Spanish dominant Mexican immigrant pupil, Juan. The videotape dialogue perhaps illustrates quite well that left to their own devices, children may use syntactically advanced structures which they might not attempt in story-telling situations--unless the forms were elicited explicitly. Such an assertion is in need of empirical validation, but it seems reasonable that when children are having a verbal argument, they are paying less attention to their language and thus speaking more naturally than when they are telling a story to other children. An example from the tape transcript:

Juan: Tienes cinco.
("You're five.")
Rosa: Siete. Si tengo cinco, tengo que estar así.
Mira. Asi.
("Seven. If I were five, I'd be like this."
Gestures with her hand.)

In the example, Rosa intends a contrary-to-fact conditional, "If I were five years old, I'd be this tall." Her Spanish utterance had present tense both for the *if* clause and for the main clause, instead of imperfect subjunctive and present conditional, as required in grammatical Spanish. Instead of using *tengo* twice, she should have used *tuviera* and *tendría*, respectively. In this instance, we have a child attempting to express certain semantic relationships for which she does not yet have the syntactic forms. This type of natural classroom discourse may well provide numerous instances which would not be tapped by a storytelling format using a sound page machine (as in this case).

●*How well suited was the BINL for teacher administration and scoring?*

In this pilot study, teachers did not abide by the BINL specification for collecting data. They indicated that since their contact with students was limited to only a few hours a week, there was not time to get the students fully acquainted with both the BINL language activity materials and the sound page machines before collecting the data. (Originally we had been informed that these schools provided half-day rather than pull-out type bilingual programs.) In fact, some of the initial "practice" sessions were unfortunately used as actual data sessions. Further, teachers sometimes chose the pictures that students were to tell stories about, rather than leaving this up to the students. Also, the teachers generally presided over the taping sessions, and one in particular hovered over the students, asking questions to elicit language samples. In fact, it appeared that the presence of the teacher as coach was more the rule than the exception. Consequently, we checked with the test developer and learned that this condition of the BINL assessment is frequently violated in the earlier grades. Thus, the BINL is perhaps not effective at eliciting so-called "natural" language from peers without adult intervention--at least not in the early grades, as in this study.

With respect to transcription and scoring, there were also problems. Teachers did not always transcribe the stories accurately, if at all, and did not always

score them properly. Perhaps the main difficulty was in separating sentences according to the BINL system in which a sentence is defined as consisting of at least an expressed or understood subject and a predicate, but possibly containing several clauses. Teachers were to take any 10 sentences in a row, and evidently had problems determining where to break sentences off. For example, the second compound clause introduced by "and" is to be eliminated altogether as a run-on. But when is a clause introduced by "and" clearly a run-on clause and when is it the beginning of a new thought? BINL stipulates an arbitrary rule based on structure rather than on meaning, namely that "and" can be allowed only once to join clauses. Thus in the sample, *Then the teacher took a stick and chased the dog and then the cat came down* (Herbert, 1977), the sentence ends after the word "dog." "And then the cat came down" is to be eliminated instead of being considered as the next sample in order to give students the opportunity to use a more complex sentence (worth more points) for the next sample. Yet, the instructions manual includes an example of a run-on sentence that is *not* eliminated, but rather scored as a separate sample, as if to suggest that such a scoring procedure *is* acceptable (e.g., "[and] I always go." [Herbert, 1977]). The teachers mostly neglected to delete a third coordinate clause. Instead, they included it in the same sentence, thus inflating the point count for that sentence.

Furthermore, teachers did not always elicit the minimum of 10 complete or partial sentences per subject. Even after three training sessions, teachers still did not write 10 sentences onto the score sheets 50 percent of the time even though it was considered acceptable according to BINL procedures to take sentences from more than one story to obtain a corpus of 10. While the BINL does allow for a corpus to have fewer than 10 sentences because some children may have very limited language skills, in this case all the children had sufficient skills in both languages to produce the necessary corpus since those who did not were eliminated from this study. Thus, the teachers had 10 sentences available for transcription.

Teachers also had difficulty transcribing the stories accurately and at times seemed to be "fixing up" what the children said, not just in terms of grammar, but in terms of vocabulary as well. This apparent "doctoring" may not have been intentional. An article by Knibbeler (1975) indicates different investigators can vary dramatically on a transcription of the same data. However, the fact that the teachers did not score some samples (one teacher did not score *any*) runs contrary to one main purpose of the BINL, mainly to get teachers close to their students' speaking data. This is why the scoring system was designed in a simplified way.

It may well be that this sample of teachers was atypical and that they were working within conditions unsuitable for the BINL--namely a pull-out program rather than an intact classroom situation. It may also be that these teachers were not particularly quick at grasping how the BINL system works. But then these teachers may reflect the common denominator of teachers that the BINL or instruments like it *must* cater to. A fairly recent review of the BINL also notes teacher difficulty in scoring the samples (Stansfield, 1978).

Generally, under the conditions described above, the BINL was not well-suited for teacher administration and scoring. It is interesting to note that subsequent to our pilot study, machine scoring became an option (Herbert, 1977) and thus somewhat reduced the teacher time required by the BINL.

DISCUSSION AND CONCLUSIONS

This study was undertaken in order to consider at close range some of the issues involved in assessing bilingual speaking skills--specifically, language proficiency, language dominance, and language change. The BINL was selected as the instrument of reference in order to facilitate the analysis, and not primarily in order to critique or evaluate it. Readers can draw their own conclusions about when and why they would use the BINL. The purpose herein is simply to illustrate how one might scrutinize or evaluate the usefulness of a specific instrument which purports to measure such concepts as proficiency, dominance, and natural language.

This involved examining the actual procedures for interpreting the results.

The BINL draws on two indices in order to measure oral language *proficiency*--average sentence length and the frequency of certain grammatical structures. In order to measure dominance, examiners must make their own decisions. The BINL manual gives no guidelines as to how this is determined. Nor does it give any statements about the comparability of structures across languages. With respect to change, there is no indication within the BINL system as to what "gain" or "loss" of certain grammatical structures might actually consist of numerically. Clearly, there are other instruments on the market which similarly lack specifics in these areas, so the problem of rigorous definitions of terms and detailed guidelines for the interpretation of results is not limited to the BINL.

As stated at the outset, the BINL was approached as a measure of natural language. It seemed appealing to build a language assessment instrument right into the normal classroom routine as one of the students' activities. It was apparently in pursuit of this goal that the scoring system was kept relatively simple, i.e. for the use of the teacher without special linguistic training. Despite the simplified level of the materials and despite teacher training sessions, it became clear, at least with respect to the Illinois pilot, that teachers might not use the BINL as its designers originally intended. For example, teachers might provide inadequate pretraining for students, intervene in the collection of stories, inaccurately transcribe and score stories, etc. The conclusion that could be drawn from this study is that an outside evaluator, preferably one with linguistic training, assist in the administration, scoring, and interpretation of an assessment instrument such as the BINL.

This study also underscored the need for the scorer to continually re-examine the raw data to make sure that the scores reflect actual language strengths and deficits rather than idiosyncrasies in the scoring system. In the case of the BINL, an examiner could look for inflated scores in grammatical categories due to the nature of the story or for sentences that are incorrectly divided, thus artificially increasing pupils' average level of complexity scores.

The issue of language parity or comparability within grammatical categories across languages was also underscored by this pilot study. It was noted that certain phenomena under investigation (such as sentence length) may differ across languages. In this case, English sentences generally seemed to be longer than Spanish sentences. Close order analysis revealed systematic subject pronoun deletion in the Spanish samples (as is acceptable in Spanish), while subject pronouns were present in English samples. This is one linguistic fact that can help to explain why Spanish sentences were shorter on the average.

Ideally, an instrument like the BINL would give indications as to what level of language behavior might be expected at different age/grade levels from mono- and bilingual children. Without such norms, it is difficult to compare students' results within the same classroom, across classrooms at a given level, and across grades. Norms for expected average language behavior--sentence length and average levels of complexity (with subcategory breakdowns)--at different grade levels and across different ethnic and socioeconomic groups would make such an instrument more meaningful. Certain norms may now have been developed for the BINL.

The results from pilot testing the BINL may simply indicate that there is no easy way to get a bilingual measure of "natural" language. Samples will vary by child and by the language the child is being asked to respond in. Other means are available for obtaining more natural language, such as that of videotaping, using wireless microphones, etc. But this approach does not begin to be feasible for large samples of bilingual children. For example, the sample videotape transcript between Rosa and Juan is almost entirely in Spanish. "Natural" language may imply limited or no data on children's speaking ability in a weaker language that they prefer *not* to speak (at all or at a given time).

Perhaps more than anything else, this study suggests that assessment in the real world--in the classroom--often results in less than ideal results. In this

case, an instrument intended to measure "natural" language may actually measure something else, namely, stories elicited by teachers (particularly with younger children). This study also points out the need to continually evaluate assessment instruments, particularly in a young field where the tests themselves are new.

REFERENCES

Cohen, A.D. "The Culver City Spanish Immersion Program: How Does Summer Recess Affect Spanish Speaking Ability?" Language Learning 24:55-68, 1974.

Herbert, C.H. Basic Inventory of Natural Language: Instructions Manual (revised edition). San Bernardino, CA: CHECpoint Systems, 1977.

Knibbeler, W. "Subjectivité dans l'Evaluation Analytique de l'Expression Orale Libre." ITL Review of Applied Linguistics 30:25-53, 1975.

Labov, W. "The Logic of Non-Standard English." The Florida FL Reporter 7(1):60-75, 169, 1969.

Matluck, J. and B. Mace-Matluck. "The MAT-SEA-CAL Instruments for Assessing Language Proficiency." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 20, 1976.

Nie, N.H. and others. Statistical Package for the Social Sciences (second edition). New York: McGraw-Hill, 1975.

Slobin, D.I. "Cognitive Prerequisites for the Development of Grammar." In Studies of Child Language Development, edited by C.A. Ferguson and D.I. Slobin. New York: Holt, Rinehart and Winston, 1973.

Stansfield, C.W. "Review of Herbert, Charles H. Basic Inventory of Natural Language." Modern Language Journal 62(1-2):64-65, 1978.

Stolz, W. and M. Bruck. "A Project to Develop a Measure of English Language Proficiency." Final Report to the National Center for Education Statistics. Arlington, VA: Center for Applied Linguistics, 1976.

Swain, M. and M. Bruck, eds. "Immersion Education for the Majority Child." The Canadian Modern Language Review 32(5), 1976.

Assessing Bilingual Reading Skills: Criterion-Referenced versus Norm-Referenced Testing of Spanish Reading

Flora V. Rodríguez-Brown
Andrew D. Cohen

Although the number and kinds of U.S. bilingual programs have increased since the inception of the bilingual education concept, the largest growth has been in Spanish-English programs. This expansion of programs has increased the need for suitable instruments to assess students' bilingual reading skills--particularly Spanish reading skills. Unfortunately, only a few instruments exist that measure Spanish reading skills, and translation of an English instrument to Spanish is not the solution given the real differences between the two reading systems (cf. Sanchez-Sadek et al, 1974). For example, the process of decoding vowels in English involves differentiating between long vowels, short vowels, and vowels with controllers (i.e. dipthongs); in Spanish, neither the position of the vowels nor the stress affects the value of the vowels. Thus, items to measure long and short vowels would not be included on a Spanish reading test as on an English reading test. Tests must be developed which take into account the basic reading processes inherent in Spanish.

Ideally, such instruments would not only measure an individual's reading performance in terms of grade level, but also yield specific feedback to teachers concerning areas in which children are experiencing difficulty or have achieved competence. Such instruments would aid educators in placing children at appropriate reading group levels as well as assist them in planning an effective reading program to meet the specific needs of individual children.

We examined and compared two different tests being used to assess such Spanish reading skills--the *Prueba de Lectura* and the SOBER-*Español* (System for Objectives-Based Evaluation of Reading-Spanish). In an attempt to determine the usefulness of each measure in a bilingual setting, we compared students' performance on, and teachers' reactions to each of the tests. We selected these instruments because each represents a different methodological approach (norm- vs. criterion-referenced) and because of the differences in format of items and types of performance each requires from subjects. The *Prueba de Lectura* is a *norm-referenced* test which assesses only reading vocabulary and comprehension skills (Guidance Testing Associates, 1967); the SOBER-*Español* is a *criterion-referenced* test which consists of: (a) "the structure of the SOBER reading domain," a framework statement which details the four main components of reading, encoding, decoding, vocabulary, and comprehension; (b) a set of behavioral instructional objectives based on the framework; (c) a guide for the users' selection of objectives; and (d) a pool of test items to assess achievement of each objective. Table 1 on the next page compares the characteristics of norm- and criterion-referenced tests (based on Housden and LeGear, 1972).

As can be seen from the characteristics in Table 1, and since there are no national norms available (although some local norms have been developed by individual school districts), we are using the term norm-referenced rather loosely in connection with the *Prueba de Lectura*. Furthermore, the publisher provide neither content analysis of what the individual items test nor any clear indication of the validity of the test. The Spanish version appears to be simply a translation of the English version, although the author claims that it is an adaptation. This version was piloted in Puerto Rico with middle-class children.

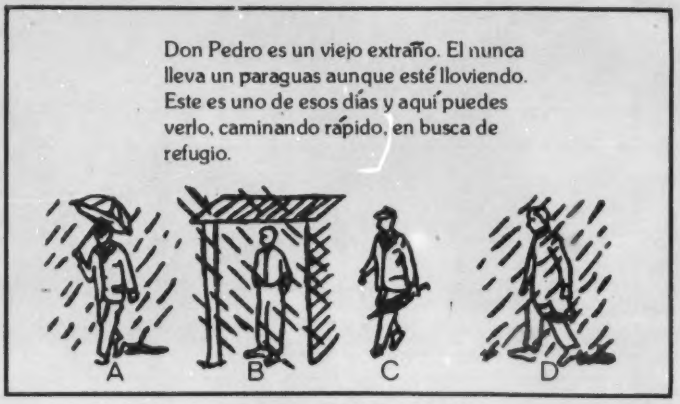
TABLE 1

Characteristics of Norm- and Criterion-Referenced Tests

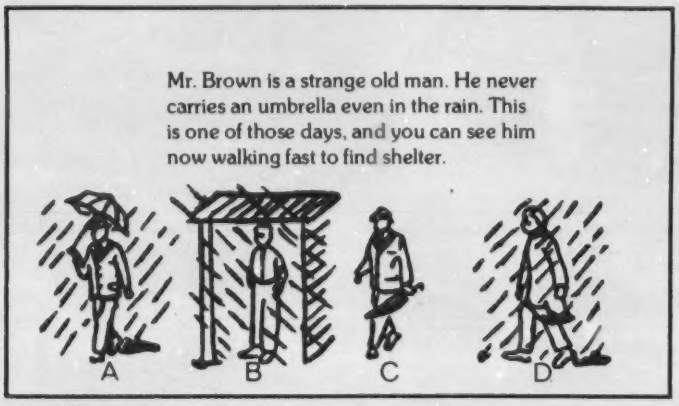
NORM-REFERENCED TESTS:	CRITERION-REFERENCED TESTS:
Evaluate individual performance in comparison to that of a group.	Evaluate individual performance in relation to a fixed standard.
Could be used to evaluate students as "below grade level level," "at grade level," or "above grade level."	Are not concerned with grade level descriptions.
Are not intended to indicate which individuals have mastered the spectrum of instructional objectives.	Identify individuals who have mastered the spectrum of instructional objectives and so apply directly to the individualization of instruction.
Are not intended to be used in planning instruction and individualization of instruction since they do not speak to specific objectives in the curriculum.	Are supposed to provide information to be used in planning instruction.
Do not operationally define mastery and/or success in specific areas.	Define mastery and/or success from an operational point of view.
Produce results that tend to be more summative than formative in nature.	Produce results that apply more to the formative than to the summative evaluation domain.
Are not very sensitive to the effects of instruction.	Are content-specific and thus sensitive to the effects of instruction.
	Are not usually standardized.
	Interpret the results in reference to a person's position in the curriculum.
	Are evaluated in relation to the objectives of the curriculum.

The SOBER-*Español* was developed in the early 1970s, and is becoming more popular due to increased emphasis on criterion-referenced tests. The test is not a translation from English; it is a comprehensive evaluation system of Spanish reading based on behavioral objectives and oriented toward the users. Items in the SOBER system reflect as many as 16 different test formats and represent different levels of difficulty within an already defined system.

Figures 1 (below) and 2 (on the following page) are samples from both tests used in the study. Figure 1 shows an example of an item in Spanish from the *Prueba de Lectura* and its counterpart item in the English reading test. Figure 2 shows a sample of an objective, the items used to measure that objective, and the instructions for administration of those particular items as they appear in the SOBER system.


Don Pedro es un viejo extraño. El nunca lleva un paraguas aunque esté lloviendo. Este es uno de esos días y aquí puedes verlo, caminando rápido, en busca de refugio.
A
B
C
D

Spanish Item


Mr. Brown is a strange old man. He never carries an umbrella even in the rain. This is one of those days, and you can see him now walking fast to find shelter.
A
B
C
D

English Item

FIGURE 1

Sample Items from the Inter-American Series--Prueba de Lectura

FIGURE 2

Sample of an Objective, Instructions, and Items from the SOBER System

Objective:

Given a written word, the learner will identify it in a set of given words.

Instructions.

Miren los cuadros con los burritos.
Miren el cuadro que tiene un burrito.
Miren la palabra que está en el cuadrito.
Miren las otras cuatro palabras.
Busquen la palabra que es la misma que la palabra que está en el cuadrito.
Llenen el círculo enfrente de esa palabra.

- ola
- oso
- oro
- ojo

Ahora miren el cuadro que tiene dos burritos.
Miren la palabra que está en el cuadrito.
Miren las otras cuatro palabras.
Busquen la palabra que es la misma que la palabra que está en el cuadrito.
Llenen el círculo enfrente de esa palabra.

- abrigo
- alma
- abeja
- ala

Ahora miren el cuadro que tiene tres burritos.
Miren la palabra que está en el cuadrito.
Miren las otras cuatro palabras.
Busquen la palabra que es la misma que la palabra que está en el cuadrito.
Llenen el círculo enfrente de esa palabra.

- escritorio
- estufa
- estrella
- escoba

Items:

1	2	3
oso	abeja	escoba
O ola O oso O oro O ojo	O abrigo O alma O abeja O ala	O escritorio O estufa O estrella O escoba

The purpose of our study was to examine and compare two different tests being used in bilingual programs to assess Spanish reading skills. Although our focus is on the *Prueba de Lectura* and the SOBER-*Español*, we feel this approach is generalizable for comparing the efficacy of norm- and criterion-referenced testing systems in terms of reliability and teacher judgments.

ASSESSING SPANISH READING SKILLS

Subjects

Our sample consisted of four separate groups (Table 2) of second graders and their teachers in bilingual programs from three school districts. Groups A_1 and A_2 came from the same school district and had the same teacher. Group A_1 attended the morning bilingual program and Group A_2 the afternoon. Group B was composed of children attending three different bilingual programs in another school district, and Group C was composed of children attending four different bilingual programs in a third district. Since one of the teachers taught in two half-day classes--morning and afternoon--a total of nine classes participated.

TABLE 2

Description of Sample

Group A_1	(N=11)	Learning to read in Spanish only
Group A_2	(N=11)	Learning to read in Spanish and English concurrently
Group B	(N=15)	Learning to read in Spanish and English concurrently
Group C	(N=27)	Learning to read in Spanish and English concurrently

N's include only students who took both the pre- and post-tests

The requirement for the project was that participating children already have at least one year of Spanish reading instruction. Eight classes were reported by the program coordinators to have been receiving reading instruction in Spanish and English concurrently, while one class was receiving instruction in Spanish. The teachers reported that about 68 percent of the children participating were Spanish dominant, 13 percent were English dominant, and 19 percent were balanced bilingual.

Eight teachers participated in the project, and all came from school districts offering bilingual programs for four or five years. Three teachers (one in district A and two in district C) taught in half-day bilingual programs (the children spent a half-day in the bilingual program and a half-day in the regular program). Four teachers (two in district B and two in district C) taught in self-contained classrooms (children attended the bilingual program all day), and one teacher (in district B) taught in a pull-out program (children were pulled out of regular classrooms to attend the bilingual classroom for a few minutes each day, for a minimum of 90 minutes per week). The average adult-student ratio was 12 children per adult.

All teachers reported that they had a teacher aide in their classrooms; all reported that they used Spanish primarily, both for Spanish language arts instruction and general classroom instruction. Most of the teachers favored the idea of first introducing reading in the native language, while the children acquired oral competence in the second language.

According to the teachers, they used different curriculum materials to teach English and Spanish reading even within the same school district. They also reported using a variety of grouping methods for teaching reading: ability groupings, individual tutorial, programmed instruction, interest groupings, and total class groups.

The teachers specified that they spoke Spanish in the classroom from 50 to 85 percent of the time (three reported 50 percent). At the same time, their estimate of the percentage of time the students spoke Spanish in the classroom ranged from 25 percent to 95 percent. The teachers, in general, tended to encourage the use of both Spanish and English in the class; some specified to the children when they expected use of each language (i.e. Spanish for language arts in Spanish, English for language arts in English, or English in the morning and Spanish in the afternoon).

Measures

The Inter-American Series *Prueba de Lectura*, Level I, was administered to the children participating in the project, Form CEs as a pretest and form DEs as a post-test. The test has 80 items--40 for *vocabulary* and 40 for *comprehension*. Figure 3 on page 29 shows examples of vocabulary and comprehension items, respectively. A "parallel" test exists in English. It is a timed test; eight minutes are allowed for the vocabulary subtest and 10 minutes for the comprehension subtest. Both the instructions and the test are in Spanish. The test is administered on a group basis.

As mentioned earlier, the SOBER-*Español* consists of four components: encoding, decoding, vocabulary, and comprehension; a list of objectives associated with the respective components; and a pool of items to measure the attainment of the objectives.

The language used in the test was controlled for levels K-3, 4-6, and 7-12, and took into account the language differences of children from different Spanish backgrounds (i.e. Mexican Americans, Puerto Ricans, etc.).

To design the SOBER-*Español* test used in this study, the eight teachers involved in the project met with the director of the SOBER-*Español* project in a five hour workshop to discuss the nature of the system and the best way to choose the objectives they would be teaching in their second grade classrooms.

Two weeks before this meeting, the teachers were given the total list of guidelines for selecting objectives for the test--objectives that they were intending to cover that year. These objectives were then discussed at the meeting, and the teachers reached a compromise on objectives for a test. The list of objectives chosen appears in Table 3 following, and a sample objective and test items that go with it appear in Figure 2 on page 25. The list of objectives was submitted to the publisher and a test was developed which had three items per objective.

The test was given by a test administrator on a group basis without a time limit and the test administrator tried to make sure most of the class had answered one particular item before going on to the next one.

Ideally, we would have liked for teachers to select their *own* set of objectives as the system is actually intended to be used. Unfortunately, economic reasons made that approach impossible. We used the group approach precisely because it was deemed a more economically feasible alternative.

Test Administration

The tests were administered by research assistants of the Illinois Bilingual Evaluation Center (IBEC) who were trained before they went to the schools. The norm-referenced test was administered in October 1975 and in April 1976. The test period was carefully timed. After instructions were given, the children had a few minutes to ask questions or to request clarifications before they started the test.

Two weeks after the norm-referenced test was administered, in October 1975 and April 1976, the criterion-referenced test, as designed by the teachers, was administered. The same test was used for pre- and post-testing purposes. The test was administered in two one-hour sessions. During each session, the children were

TABLE 3

List of Objectives for the SOBER Test Used for the Project

1109 Given two words orally, the learner will indicate whether the words rhyme.

1133 Given a set of words orally that have the same medial vowel sound, the learner will produce the medial vowel sound orally.

1154 Given a manuscript uppercase (or lowercase) letter, the learner will write its corresponding lowercase (or uppercase) form.

1202 Given a word orally that has a single initial consonant, the learner will write the initial consonant.

1222 Given a word orally that has an initial consonant blend, the learner will write the initial consonant blend.

1242 Given a word orally that has a single vowel, the learner will write the vowel.

1262 Given a word orally that has a dipthong, the learner will write the dipthong.

1301 Given a word orally, the learner will write it correctly.

2102 Given a written word, the learner will identify it in a set of given words.

2201 Given a set of manuscript uppercase and lowercase letters and given the name of one of the letters orally, the learner will identify the letter named.

2202 Given a word orally and a set of numerals in print, the learner will identify the numeral that represents the number of syllables the word contains.

2203 Given a set of written words, the learner will identify the word that contains an initial consonant blend.

2222 Given a written letter, the learner will identify from a set of written letters the one that precedes it in the alphabet.

2261 Given a written consonant, the learner will pronounce the sound the consonant represents.

2401 Given a written consonant digraph, the learner will pronounce the sound the digraph represents.

2501 Given a written vowel, the learner will pronounce the sound the vowel represents.

3105 Given an illustration of a word and a set set of written words, the learner will identify the word that corresponds to the illustration.

3122 Given a written sentence and a set of illustrations, the learner will identify the illustration of the sentence.

3203 Given an incomplete written sentence and a set of written words, the learner will identify the word whose meaning is appropriate to the context of the sentence.

3323 Given a written sentence with the verb missing and a set of written inflected forms of a verb, the learner will identify the form that correctly completes the sentence.

3382 Given a written feminine word and a set of written words, the learner will identify the masculine counterpart of the feminine word.

3401 Given a written word and a set of written definitions, the learner will identify the definition of the word.

4101 Given an oral selection and a set of illustrations, the learner will identify the title that best expresses the main idea of the selection.

4103 Given an oral selection and a set of written titles, the learner will identify the title that best expresses the main idea of the selection.

4201 Given an oral selection from which the main idea can be inferred and given a set of illustrations, the learner will identify the illustration that best represents the main idea.

4301 Given a written selection and a set of illustrations, the learner will identify the illustration that best represents the main idea of the selection.

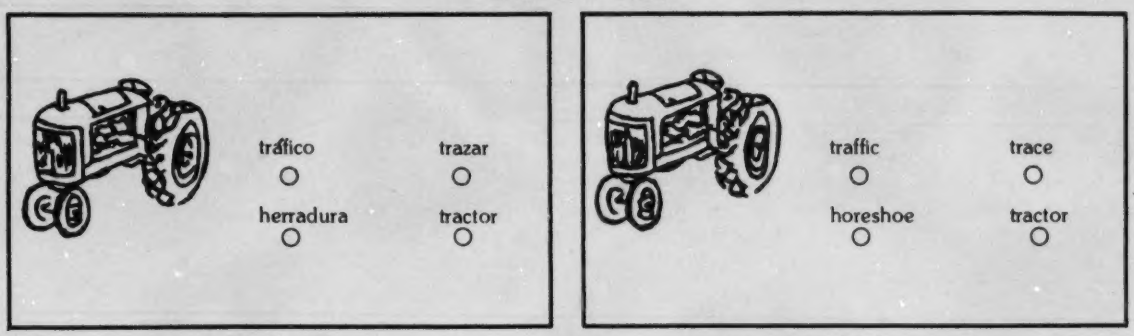
4401 Given a written selection from which the main idea can be inferred and given a set of illustrations, the learner will identify the illustration that best represents the main idea of the selection.

4461 Given a written selection in which the conclusion is implied and given a set of written sentences, the learner will identify the sentence that best expresses the conclusion.

FIGURE 3

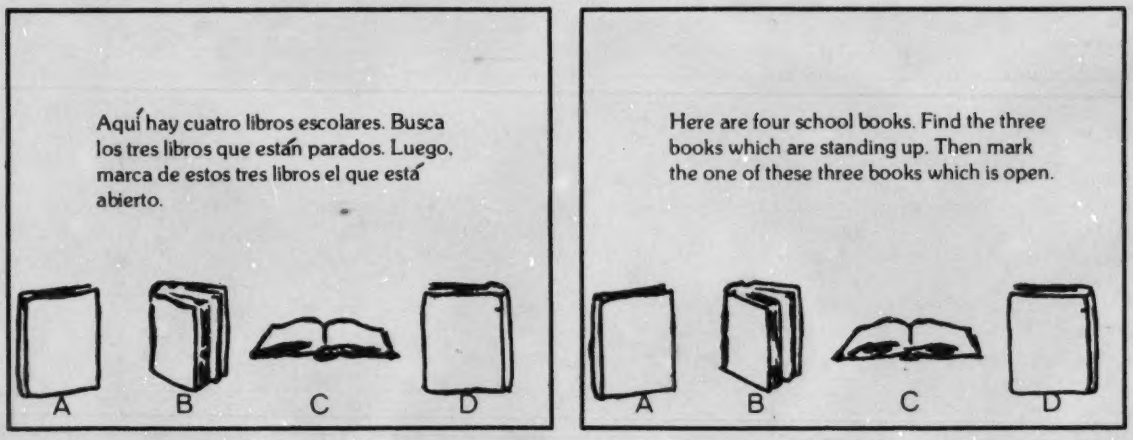
Sample of Vocabulary and Comprehension Items from the Prueba de Lectura

VOCABULARY ITEM



Spanish English

COMPREHENSION ITEM


Aquí hay cuatro libros escolares. Busca los tres libros que están parados. Luego, marca de estos tres libros el que está abierto.
A
B
C
D
Here are four school books. Find the three books which are standing up. Then mark the one of these three books which is open.
A
B
C
D

Spanish English

given a short break after a half-hour of testing. The test was administered to intact classroom groups, except for three items which required individual administration.

Data Analysis

Tests were hand scored. Mean percentages of correct responses, standard deviations, and standard error of means per school district as well as for the total sample were calculated for the two tests. The SPSS (Nie et al, 1975) system was used to calculate Pearson correlation coefficients between the two tests by subtest (vocabulary and comprehension) for the total sample and for each school district independently.

Pre- and post-test differences in mean percent correct on the vocabulary and comprehension items were calculated by means of T-tests. Kuder-Richardson 20 reliability coefficients were calculated for both the vocabulary and the comprehension items on post-test scores of the two tests. Teachers received their students' pretest results after the tests were scored. The teachers were also provided with an explanation of what the scores meant and how these results could be applied in the classroom.

FINDINGS

•*In what ways did the objectives for the criterion-referenced test selected by teachers as being "essential" to their Spanish reading program match objectives in district guidelines, curriculum materials guides, and teachers' own lesson plans?*

Since the eight teachers selected reading objectives in consort, they endorsed objectives which they otherwise may not have chosen because they were too advanced or too easy for their specific groups. While three teachers complained that some objectives were hard for their students, a teacher from a different school district complained that some items were too *easy*.

Nevertheless, most teachers felt that the objectives selected for their criterion-referenced test were essential and matched the objectives of their Spanish reading programs. One teacher expressed the view that children who were learning to read in Spanish and English concurrently were at a disadvantage in both languages since teachers could not teach *both* a full Spanish and English reading curriculum (as required by individual districts at a given grade level) in the same period of time as they could teach reading in one language.

•*What were the in-depth reactions of teachers toward a set of objectives selected from a criterion-referenced system and toward the corresponding test items in comparison with their reactions towards other means of assessment (e.g., teacher-made tests, tests in curriculum guides, and standardized tests)?...How well did the teachers feel the items in their criterion-referenced test were measuring their respective objectives?...Does the criterion-referenced test adequately assess students' Spanish reading skills at the beginning of grade two?*

Before requesting that teachers compare criterion-referenced tests to other tests, they were asked if they knew of other tests, not necessarily reading tests. Five of the eight teachers knew of at least one test. They named, among others, the *Boehm Test*, the *Peabody Picture Vocabulary Test*, the *Metropolitan Readiness Test*, and the *Test of Basic Experience*. Teachers were asked to specify prior ways in which Spanish reading had been assessed in their schools. They reported that Spanish reading had been assessed either by means of teacher-made tests of reading comprehension or by informal teacher observation.

Teachers thought that the procedure of selecting the criterion-referenced test objectives was good and that the objectives *did* reflect the school districts' expectations of what they wanted children to know at a given grade level (in this case, grade two). Here again, teachers were uncomfortable at having to choose objectives across school districts. They also felt that they should have been given more time to decide on the objectives they would teach.

According to the teachers in the study, the encoding and decoding skills were actually more relevant than vocabulary and comprehension skills, given the level of reading of the children in the project. Teachers explained that they emphasized these two aspects, encoding and decoding, in their instruction more than vocabulary and comprehension, and we note that these two areas were not even tested by the norm-referenced test.

Some teachers specified areas in their reading curriculum which were *not* covered in the test (i.e. sentence construction and punctuation), as well as others which did not pertain to the reading domain (such as creative writing and creative skills). The teachers would have liked to see more sentence-picture association questions (where a sentence or paragraph is read orally by the test administrator and the child identifies the picture that corresponds to the sentence or paragraph). Teachers from one school district recommended that the items regarding the use of medial vowels and names of letters be omitted from their test as these concepts were not taught in their programs.

In comparison to methods by which English reading was assessed in their schools, the teachers thought the criterion-referenced test used in this study was a good instrument for assessing Spanish reading. They thought it was more valid than teacher-made tests. They particularly liked the general organization of the test, its relationship to specific objectives, the developmental sequence followed in the design of the system (i.e. the grading of material according to what a child is able to do at a certain stage of development), and the variety of skills it assessed in comparison with other tests.

Teachers also thought very highly of the criterion-referenced test's physical layout and the procedures for administration. They thought that the instructions were comprehensible for the children, and that the test was not overly fatiguing to take due to the diversified ways in which questions were asked.

They thought the format was likely to be more interesting and appealing to children than that of other tests. Teachers believed that the fact that the items in the criterion-referenced test used were *not* a translation from an English version of the test (as was the norm-referenced test) made it a better test than tests that are just translations of original ones.

In general, teachers thought that the test was good at pinpointing their students' reading problems. They specified that the results were especially useful since the test measured very specific student behaviors. Four teachers remarked that the test developed from the criterion-referenced system for this study was the best test for evaluating Spanish reading in second grade they had seen up to then.

In comparing criterion-referenced to norm-referenced testing, teachers expressed the belief that standardized tests were good when conditions were controlled well and directions were given clearly, but they made it clear that these tests did not measure their specific classroom objectives. According to the teachers, standardized tests were generally not culturally fair since they are prepared for and normed on middle-class children who are not similar to typical bilingual children with respect to experiences. Teachers felt that the criterion-referenced test used was not biased against children from different ethnic backgrounds, and that it was based on more than just *adult* expectations of what children should know.

They strongly recommended the use of criterion- over norm-referenced tests to measure Spanish reading skills. One teacher, however, preferred *teacher-made* tests to commercially produced criterion- *and* norm-referenced tests since, according to her, they were more sensitive to individual needs and development.

Overall, it was the teachers' feeling that their selected criterion-referenced test measured the objectives and specific behaviors that they taught at the second grade level. In regard to tests results, teachers felt more confident about the results of the criterion-referenced test than about those from the norm-referenced test. They specified that they could use the criterion-referenced test results more effectively in their planning of activities.

In general, they believed that the criterion-referenced test could be used

TABLE 4

Correlations between the Criterion- and the Norm-Referenced Pretest Scores

District	N	Vocabulary	Comprehension
A_1 & A_2	32	.46**	.57**
B	23	.76**	.63**
C	35	.76**	.32
Overall	90	.71**	.54**

**$p < .01$

TABLE 5

Correlations between the Criterion- and the Norm-Referenced Post-test Scores

District	N	Vocabulary	Comprehension
A_1 & A_2	30	.30	.49**
B	23	.80**	.73**
C	33	.61**	.53**
Overall	86	.59**	.62**

**$p < .01$

TABLE 6

Mean Percent Correct Scores in Vocabulary and Comprehension (N=64)

	Pretest	Post-test
Norm-Referenced		
Vocabulary Subtest	.51	.78
Comprehension Subtest	.33	.58
Criterion-Referenced		
Vocabulary Items	.54	.71
Comprehension Items	.54	.61

as a valid diagnostic baseline tool to assess children's skills at entry into a particular grade. Again, teachers believed that the instrument would be more valid if it reflected the objectives for a given grade level in one particular school district and not for a group of districts together (hence the need for more capital outlay). The teachers asserted the need to receive the pretest scores and the instructions on how to use these scores in a diagnostic fashion early in the year. Seven teachers specified that they received enough information (as described before) to diagnose the overall level of each student tested. They used the information in different ways; some worked individually with students in the problem areas determined by the test results while others constructed lessons to emphasize the weak points shown by most children in the class, and still others developed their second grade curriculum using the results of the test as a baseline measure in the planning.

•*How did the pre- and post-test scores of the criterion-referenced test compare with those from the norm-referenced test used?...Did the criterion-referenced test assess gains in Spanish reading over the year?*

The cross-district pretest correlations between the two tests were significant both for vocabulary and for comprehension, with the correlation for vocabulary being higher (r=.71 vs. r=.54, $p<.01$). Post-test correlations were again significant and more similar to each other (vocabulary, r=.59; comprehension, r=.62, $p<.01$) (see Tables 4 and 5 on page 32). The mean percentage of correct responses for the overall group shows an improvement trend from pre- to post-test for both tests (see Table 6).

Differences between the mean percent correct for vocabulary and comprehension items on the two tests, both in pre- and post-testing, were computed to see whether the two tests were measuring similar vocabulary and comprehension skills. Results from the pretest indicated that in all districts the mean percent correct for the comprehension items differed significantly between the two tests (see Table 7). The students tended to do better on the criterion-referenced items than on the norm-referenced items. Mean differences on vocabulary items were nonsignificant in all but district C, where students performed better on the criterion-referenced test.

TABLE 7

Differences in Mean Percentage of Correct Responses between Pretest Items for the Norm-Referenced and the Criterion-Referenced Test

Groups	N		Mean Difference*	Standard Deviation	Standard Error	2-Tail Probability
A_1/A_2	32	Vocabulary	.64	25.14	4.44	n.s.
		Comprehension	-19.27	18.25	3.22	.01
B	23	Vocabulary	.02	17.75	3.70	n.s.
		Comprehension	-20.54	14.41	3.00	.001
C	35	Vocabulary	-8.18	15.11	2.55	.01
		Comprehension	-20.64	19.25	3.25	.001
Overall	90	Vocabulary	-2.95	20.07	2.12	n.s.
		Comprehension	-20.13	17.59	1.85	.001

n.s.=nonsignificant

**A positive difference is in favor of the norm-referenced test and a negative difference is in favor of the criterion-referenced test.*

The results on the post-test were different from those of the pretest data. The differences in mean performance on comprehension items were significant ($p < .05$) only in district A in favor of the criterion-referenced items. Differences in vocabulary items were significant in districts A and B, but not in C (see Table 8 following). The overall performance on vocabulary items seemed to improve greatly from pre- to post-test on the norm-referenced test in comparison to the criterion-referenced test.

The pretest comprehension results could be viewed as a consequence of the different strategies that both tests use to measure comprehension skills. While the norm-referenced test required the reading of a paragraph followed by a question with four alternative responses, the criterion-referenced test employed a variety of strategies, such as the identification of the title for a story read by the test administrator or the identification of the picture that relates to a story read to the child. Perhaps the comprehension items on the criterion-referenced test were somewhat easier because children did *not* always have to read something in order to answer them, which *was* the case with regard to the norm-referenced items.

As for the post-test comprehension results, by the time of the post-test perhaps the children had mastered the strategy used in the norm-referenced test as well as in some of the criterion-referenced test items, so that the difficulty of both sets of comprehension items was equalized. Since the norm-referenced test is a timed test, it is possible that children were able to read faster and thus performed better on the post-test. With respect to pretest vocabulary performance, vocabulary items from the norm-referenced test required the children to read a word and find a drawing of it among several pictures presented. In the criterion-referenced test there was more variety of strategies for testing vocabulary, such as giving the children a picture whereby they identified the word for it or giving the children an incomplete sentence with four alternatives whereby they identified the right one. Although the tests used different strategies for vocabulary, both tests required the children to know or recognize a word. If the word was not known, children had the same chance of marking the right answer in both tests. This may account for the nonsignificant differences in pretesting. The results from the post-test seem to indicate that children did much better in the vocabulary section of the norm-referenced test. This may be due to the fact that children had learned more of the vocabulary included in the norm- than in the criterion-referenced test or that the norm-referenced test was more sensitive to change. In absence of a separate test of knowledge of vocabulary on both tests this explanation remains speculative.

The mean percent correct on vocabulary and comprehension items for the two tests for the pre- and post-test, respectively (see Table 6) was submitted to gain score analysis. The gain scores on the two tests for the comprehension and vocabulary items appear in Tables 9 and 10, following. Gain scores for comprehension items on both tests were significant for the overall groups (Table 9). However, in two of the three school districts the criterion-referenced test gain scores on the reading comprehension items were not significant. The vocabulary gain scores for the overall sample and for school districts individually were significant in all but one instance. The criterion-referenced test vocabulary gain score was not significant in district B.

Whereas gain scores by district were consistently significant on the norm-referenced test, the criterion-referenced test gain scores by district were only significant half the time. This finding seems to show that the criterion-referenced test used in the study is not as effective as the norm-referenced test at measuring improvement or gain in reading skills over a school year.

The different formats of the two tests may provide a partial explanation for this finding. While the norm-referenced test used only one format strategy for these items (read a word or a paragraph and choose the right picture), the criterion-referenced items required one of 16 different strategies (i.e. reading a paragraph and choosing a picture related to the paragraph or choosing a written answer, or a title; writing the letter named by the test administrator, etc.).

TABLE 8

Differences in Mean Percentage of Correct Responses between Post-test Items for the Norm- and Criterion-Referenced Tests

District	N		Mean Difference*	Standard Deviation	Standard Error	2-Tail Probability
A_1/A_2	30	Vocabulary	9.23	22.78	4.16	.05
		Comprehension	-7.97	20.24	3.69	.05
B	23	Vocabulary	9.59	13.24	2.76	.01
		Comprehension	.91	16.08	3.35	n.s.
C	33	Vocabulary	3.70	20.06	3.49	n.s.
		Comprehension	-1.99	16.18	2.82	n.s.
Overall	86	Vocabulary	7.20	19.54	2.11	.001
		Comprehension	-3.30	17.84	1.92	n.s.

**A positive difference is in favor of the norm-referenced test and a negative difference is in favor of the criterion-referenced test.*

n.s.=nonsignificant

TABLE 9

Gain Scores for Comprehension Items on the Norm-Referenced (NRT) and the Criterion-Referenced (CRT) Tests

District	N	Test	Number Of Items	Mean Difference For Raw Score	Standard Deviation	Standard Error	2-Tail Probability
A_1/A_2	22	NRT	(40)	-10.82	6.77	1.44	.001
		CRT	(24)	3.88	3.88	.83	.001
B	15	NRT	(40)	-11.47	7.75	2.00	.001
		CRT	(24)	- 1.33	3.42	.88	n.s.
C	27	NRT	(40)	- 9.04	8.96	1.72	.001
		CRT	(24)	- .78	3.18	.61	n.s.
Overall	64*	NRT	(40)**	-10.22	7.927	.99	.001
		CRT	(24)	- 1.91	3.676	.46	.001

TABLE 10

Gain Scores for Vocabulary Items on the Norm-Referenced (NRT) and the Criterion-Referenced (CRT) Tests

District	N	Test	Number Of Items	Mean Difference For Raw Score	Standard Deviation	Standard Error	2-Tail Probability
A_1/A_2	22	NRT	(40)	-10.59	10.92	2.33	.001
		CRT	(24)	- 2.77	3.68	.89	.01
B	15	NRT	(40)	- 6.87	8.98	2.32	.01
		CRT	(24)	- 1.93	3.94	1.02	n.s.
C	27	NRT	(40)	-14.26	8.54	1.64	.001
		CRT	(24)	- 3.78	3.68	.71	.001
Overall	64*	NRT	(40)**	-11.27	9.82	1.23	.001
		CRT	(24)	- 3.00	3.75	.47	.001

**For both analyses only 64 subjects were used. They represent students who took the pre- and post-tests.*

***Note that SOBER-Español has a smaller total score for both comprehension and vocabulary (24 items each) than the Prueba de Lectura, which has 40 items for each subtest.*

n.s.=nonsignificant

In other words, it may have been more difficult to respond to the many different formats used in the criterion-referenced test than to the single one that was used in the norm-referenced one.

Responding correctly to different formats requires cognitive flexibility, as well as knowledge of the basic concepts that underlie the different strategies involved in finding the right answer. Its varying formats may thus make such a criterion-referenced test harder than this norm-referenced test. It is also possible that the diversity of formats is particularly distracting for younger readers such as the second graders in this study.

It is our view that the variety of criterion-referenced test formats could give the teacher a broader picture of the learners' understanding of reading tasks (in this case, comprehension and vocabulary) than the single format used in the norm-referenced test. By reacting to only one strategy (in the norm-referenced test) the subjects may answer the questions mechanically, because once the strategy is learned, not much cognitive energy is needed to answer the rest of the questions. Children with limited cognitive flexibility and short attention spans may have particular difficulty on a test with multiple formats. In that the criterion-referenced test is measuring the ability to switch sets, it is also measuring the ability to follow instructions. This may also explain why greater gains were measured over time using the single format test.

●*Were teachers comfortable with the criterion for assessing the attainment of an objective, namely, that the students answer correctly two items out of three? ...Were there any patterns shown in the answers students gave to the three items for each objective which would indicate that they were reacting to a different stimulus from the one they were supposed to react to?*

Although teachers liked the criterion-referenced test and believed it was a good instrument to assess Spanish reading skills, they thought that the two out of three criterion was not a good one. Some teachers recommended that more items per objective be added to the test.

To make sure that children were answering the questions as expected and were not reacting to something different than what the items were supposed to measure, item analysis of response patterns on the childrens' answers was performed. The analysis showed that when children had one out of the three items right, this was often a result of guessing. For example, children would mark the first alternative for each of the three items, and by chance one of the right answers was in that position. Sometimes children would choose the second alternative in each of the three items and the same situation would occur. It is important to point out that this strategy was not used throughout the whole test by any given child, but only in reaction to specific objectives that the child did not know how to answer.

Response patterns were also analyzed in instances where children correctly answered two out of three items for an objective. Except in one occasion, item analysis showed that the incorrectly answered item was not the same across all children. In the one instance where children did get the same item wrong, the children heard the name of a letter and then were supposed to write it. Children that had one item wrong for this objective wrote the letters *N* and *R* correctly, but not the letter *K*. It seems possible that the children were reacting to the sound *CA* (the name for *K* in Spanish, which is much more common in Spanish) rather than the letter *K* which is seldom used in Spanish. Indeed, the most common answer given by children was *CA*. In this particular case the children were not reacting to the right stimulus. In the case of *N* and *R*, the names of the letters were familiar to children since these letters are commonly used in Spanish.

In answer to the question about item response patterns, we can say that there were no generalizable patterns followed by children when answering the criterion-referenced test items. Each pattern analyzed seemed to be related to particular cases which appeared in this specific test.

●*What were the reliability coefficients for vocabulary and comprehension items on the criterion-referenced and the norm-referenced tests?...Overall, how did the two tests compare in terms of reliability?*

The reliability by district and overall for both the vocabulary and comprehension items on the norm-referenced test was high (over .90, see Table 11) when compared to the reliability of the criterion-referenced test items, which reached only .77 for comprehension items in district B (Table 12).

The finding that the SOBER-*Español* test had lower reliability coefficients may be partially explained by some general factors which influence the reliability of a test. These are the length of the test, the nature of the test, and the testing procedure. Each of these factors may have influenced our results.

The length of the test. It is understood that the longer the test, the greater its reliability. In the case of the criterion-referenced test, reliability was computed only on the basis of 42 items from the 90-item test. These 42 items corresponded to six objectives testing vocabulary and eight objectives testing comprehension.

Since the criterion-referenced test measures achievement in terms of objectives attained and each objective is measured by means of three items, only 14 cluster scores (one score per objective, minimum=zero, maximum=three) were tallied and used in computing the reliability of the criterion-referenced test, while item scores (one or zero) for the 40 vocabulary and 40 comprehension items were used to calculate the reliability of the norm-referenced test. It can be seen that the larger number of item scores in the norm-referenced test would naturally increase its reliability.

The nature of the test. First of all, it is important to stress that two different types of tests were being compared: a criterion-referenced test and a norm-referenced test. Further, it is necessary to emphasize again that we are comparing only part of the criterion-referenced test (14 objectives out of a total of 30 objectives) with the two regular intact subtests of the norm-referenced test.

Another difference, as explained before, is that while the norm-referenced test uses basically a single format calling only for a few response strategies, the criterion-referenced items are in 16 formats requiring a variety of strategies. Since each item in the criterion-referenced test may require something different from the students, there is no way they can memorize or practice any strategy in particular, and this may affect the internal consistency of item responses, which is what the Kuder-Richardson coefficient assesses.

The testing procedure. While the norm-referenced is a group-administered and timed test, the criterion-referenced test was not timed and some items required individual administration. The criterion-referenced test requires the child to work at the group's pace and the test administrator tries to make sure that most of the class have answered one particular item before going on to the next. In this study, the 90 items in the criterion-referenced test were administered in two one-hour sessions, giving opportunity for all children to answer all the items, so that some of the items that were used for comparison were given in the first session and others in the second session. This testing procedure may have helped performance of children with low attention spans and may have influenced the reliability of the test.

Another difference in the testing procedure is that the criterion-referenced test has special instructions accompanying each set of three items for a particular objective.

DISCUSSION AND CONCLUSIONS

The main purposes in conducting this project were: (a) To use a criterion-referenced test developed from local objectives with a small group of teachers to examine the feasibility of using the system as an alternative to norm-referenced tests previously used. (b) To compare the usefulness and applicability of the criterion-referenced test used with a norm-referenced test commonly administered in bilingual programs.

TABLE 11

Reliability Coefficients for the Norm-Referenced Test*
(Based on Post-test)

Group	N	Vocabulary	Comprehension
A_1 & A_2	34	.97	.96
B	28	.95	.94
C	42	.95	.91
Overall	104	.96	.94

**Kuder-Richardson 20*

TABLE 12

Reliability Coefficients for the Criterion-Referenced Test*
(Based on Post-test)

Group	N	Vocabulary	Comprehension
A_1 & A_2	30	.65	.75
B	23	.56	.77
C	23	.61	.20
Overall	86	.61	.74

**Kuder-Richardson 20*

The criterion-referenced test used in the study was seen by teachers and by the authors as a good alternative to norm-referenced testing. Teachers seemed to prefer it because (a) it related directly to what they were teaching in their classrooms, (b) it tested a variety of learning strategies for reading, (c) it was designed by levels, and (d) it was sensitive to developmental as well as cultural differences. In that the criterion-referenced system covers a wide domain of skills and is objective based, the teachers thought that it could be used as a diagnostic instrument for grouping children or for individual instruction.

Empirical analysis revealed that the criterion-referenced test did *not* assess gain in reading skills as well as the norm-referenced test used. Item analysis of responses to test items suggested that children were, in fact, reacting to the appropriate stimuli rather than to extraneous factors. The study shows, though, that children with limited cognitive flexibility and limited attention spans may have particular difficulties on a test with multiple formats.

Although some comparisons have been made here with respect to the reliability of the criterion-referenced and the norm-referenced tests used, the authors believe that it is difficult to compare the reliability of tests which differ in many crucial respects including objectives, length, item formats, and the conception of the reading domain. The data obtained from this small project show that the norm-referenced test used *(Prueba de Lectura)* seems to be more reliable and a better predictor of gains in students learning reading than the criterion-referenced test used, which was developed from objectives and items from the SOBER system. These results contrast with the teachers' views regarding the two methods of testing. Although teachers feel more secure when the test measures the objectives they are teaching, the criterion-referenced test used in this study did not provide the best results statistically.

Test results notwithstanding, however, the criterion-referenced instruments may prove more helpful than norm-referenced. Criterion-referenced tests act as a diagnostic tool in assessing children's reading skills. Once an accurate diagnosis is made, teachers may then make educated guesses regarding what skills to be taught and tested. Ability to pinpoint skill areas which need development seems to be a crucial factor teachers consider when looking for tests. This fact makes criterion-referenced testing more appealing to teachers and enlightens their view of the applicability of testing results to the classroom.

REFERENCES

Guidance Testing Associates. Manual and Technical Report, Tests of General Ability and Tests of Reading, Inter-American Series Forms CE and DE, CEs and DEs. Austin, TX: Author, 1967.

Housden, J.L. and L. LeGear. "An Emerging Model: Criterion-Referenced Evaluation." Thrust 2:5, 1972.

Nie, N.H. and others. Statistical Package for the Social Sciences (second edition). New York: McGraw-Hill, 1975.

Sanchez-Sadek, C. and others. "The SOBER System." Paper presented at the annual meeting of the American Educational Research Association, Chicago, May 1974.

Assessing Language Use in Bilingual Classrooms: An Ethnographic Analysis

Margaret Bruck
Jeffrey Shultz
Flora V. Rodríguez-Brown

In the past decade there have been a number of evaluations of bilingual education programs (e.g., Barik and Swain, 1975; Bruck, Lambert and Tucker, 1974; Cohen, 1975; Dillon, 1971; Goodman and Stern, 1971). While this number continues to grow most of the evaluations are product (outcome) rather than process (dynamically) oriented. Thus there is a fair amount of information documenting the effects of bilingual education programs on the academic, social, linguistic, and cognitive development of its students, but there is little information available on the dynamics of the situations in which the concerned students are involved.

Process data are crucial for the interpretation of product data. Product data yield information on how well a specific group does on a specific test which purports to measure some skill or ability. Process data yield information on why the group performs the way it does on that test. Process evaluation is concerned with the description of events in the classroom which are the educational causes of the effects measured by product evaluations. Ideally the two types of evaluations would be carried out in conjunction with each other; first however, a method for undertaking process evaluations must be developed. The purpose of this pilot study was to examine the educational environment of bilingually schooled children by means of techniques derived from the work of ethnographers of speaking (cf., Erikson and Shultz, 1976; Hymes, 1976). Specifically, the patterns of language used in the environment of two first grade bilingually schooled children (Spanish/English) were analyzed from videotape data three times during the school year to systematically describe the process or nature of events that occur in a bilingual classroom.

This study was primarily undertaken for applied reasons. It was a component of a large-scale evaluation of the bilingual programs in downstate Illinois. This component represented an attempt to study some of the processes of bilingual education, using techniques that others have used to study social interaction in Anglo classrooms (cf., Mehan et al, 1976; McDermott, 1976; Shultz, 1976).

Results of the present study are of relevance for pedagogical, theoretical, and methodological reasons. Pedagogically, there appear to be few process evaluations which are educationally relevant. In a review of Title VII evaluations, a few process evaluations were found which used a variety of methods to describe the dynamics of a specific program or classroom. For example, teachers were given checklists and asked to indicate the percentage of time devoted to different activities (Cox, 1975). In another program, teachers were administered the Flanders Interaction Analysis Scale three times during the year to determine the amount of teacher talk-time (Corpus Christi, 1975). In other programs, individual students or classrooms were observed for selected periods during the school year to determine the amount of language used for different activities and situations (Chula Vista, 1973; Austin, 1975; Cohen, 1975). In one evaluation (Beacon, 1973), on the basis of classroom observations, the evaluator rated 19 instructional processes related to bilingual education on a five point scale (e.g., students responded to teacher in the same language addressed; teacher transmitted the attitude "the school cares" to the student). The greatest shortcoming of many of these evaluations is the lack of generalizable statements about the processes of a bilingual education program.

In order for these process evaluations to be of significance for the practitioner, some general conclusions about what is going on in the bilingual classroom must be drawn. Educators need to know: *How bilingual is a bilingual classroom?* Are the models that are proposed for specific programs (e.g., alternate days, concurrent, half-day pull-out) actually being followed in the classroom by the teacher and pupils? Detailed information is required concerning the extent to which each language is used, as well as the conditions under which each language is used by the participants. Our study addressed some of these issues on an exploratory basis.

The results of this investigation provide some data relevant to the study of the ethnography of speaking. This is the study of what a participant must know to act appropriately in a given culture (Goodenough, 1964). It is a sociolinguistic fact that every speaker controls a number of varieties of speech patterns. The choice of a specific pattern is not random, but determined by a number of contextual constraints: where the speakers are, who is present, the topic, the relative status of the speakers, etc. The purpose of an ethnography of speaking is to explain these patterns of language use in terms of the sociocultural meanings associated with each variety. Thus, an ethnography of speaking not only yields information on when and how different speech varieties are used, but also makes inferences about the social and cultural meanings which are involved in selecting one form as opposed to another (cf., Blom and Gumperz, 1972).

This kind of study relies on naturalistically collected data. In order to understand the patterns of language use in a specific community, detailed data on the contexts in which the language is used must be known (i.e. the topic of conversation, the relative status of the speakers, where the interaction occurs, etc.). This can only be done if language samples of actual social interaction are collected.

There have been several studies of social interaction in bilingual classrooms. Genishi (1976), Legarreta (1977), Phillips (1975), and Shultz (1975) each examined the amount of Spanish and English used in bilingual classrooms as well as the conditions under which both languages were used. Genishi found that although Spanish was used more often than English in structured classroom situations, the children chose to use English more frequently in unstructured situations. Phillips, Legarreta, and Shultz all found a predominant use of English in the classrooms. It appears that English is used for discipline and control while Spanish is used primarily for instruction. Results of these studies strongly suggest that English is the dominant language in the classroom and that the children's vernacular is of relative unimportance in the educational setting.

While all the above studies were carried out in bilingual classrooms, the researchers were primarily interested in the phenomenon of language use rather than the processes of bilingual education per se. Thus, the data analysis as well as the types of conclusions drawn are not as pertinent to practitioners as they might be. This study attempts to address both theoretical and practical issues related to patterns of language use in a bilingual classroom.

Methodologically, this study makes two contributions. First, to our knowledge, it is the first bilingual classroom study to use videotape equipment. All previous studies have either used questionnaires, on-site coding procedures, or audiotape procedures. Questionnaires and self-reports are highly unreliable especially in the study of language use patterns (Blom and Gumperz, 1972). On-site coding is also unreliable: the researcher has no recourse to his observations and most make on-the-spot decisions which are often highly subjective (e.g., uses language "A" to express warmth) or difficult to code for the first time (e.g., it is often difficult to determine which language is being used at a particular time especially if mixing of languages occurs). While the use of audiotape solves some of these problems, one still loses much detailed information on the nonverbal contexts of the interactions which are crucial for their understanding. The use of video equipment solves these problems: on-site judgments need not be made, and one has a visual record of the contexts of the interactions that can be played back to check on the reliability of rater's judgments.

The second methodological contribution involves the attempt to use video and ethnographic methods for program evaluation. While we believe these methods to be extremely suitable for the study of classroom dynamics, it may be that these techniques are too expensive in terms of time and equipment, too sophisticated for the untrained researcher, and too limiting in terms of the number of situations and events that can be studied for the evaluator. Our study deliberately tried to address these questions and problems by training research assistants to do the data collections and preliminary data analysis. Careful records were kept of the amount of time invested in the project as well as the numbers and types of problems that occurred.

Thus, the purposes of the present pilot study were twofold: to better understand the process of education of bilingually schooled children, and to investigate the feasibility of using videotape and ethnographic procedures to examine this process in an evaluation context. Specifically, ethnographic procedures might be used generally to obtain the following information:

- *How much each language (Spanish/English) is used during teacher-child interactions in bilingual classrooms; during child-child interactions in bilingual classrooms; during child-child interactions in nonbilingual classrooms.*
- *Is there a difference in the amount of Spanish used during peer-peer interaction in bilingual and nonbilingual classrooms?*
- *Does the relative amount of language used in a specific context change as a function of time (over the school year) and/or as a function of teacher's language proficiency?*
- *Why and under what conditions each language is used during teacher-child interactions in bilingual classrooms; during child-child interactions in bilingual classrooms; during child-child interactions in nonbilingual classrooms.*
- *Why code switching (i.e. a change of language during a conversation) occurs; do the rules for language use in specific settings change as a function of time (over the school year) and/or as a function of teacher's language proficiency?*

AN ETHNOGRAPHIC ANALYSIS

Subjects

Students attended bilingual classes for half-day and regular English classes for half-day. The bilingual and all-English schools were situated in a Chicago suburb. The bilingual program was characterized as transitional half-day pull-out. Children were bused from neighboring schools to attend the program either in the morning or afternoon. They were selected by their "home" school to attend the program because their knowledge of English was insufficient to cope with the demands of English classrooms. Once their English language proficiency became adequate to meet these demands, participation in the bilingual program would be terminated (transitional).

Two children, Maria and Rosa, were selected to serve as target subjects. Both attended the same bilingual class. Maria attended in the morning and Rosa attended in the afternoon. The children, both female, attended two different English schools. In each of their regular classes there were children of similar linguistic backgrounds as well as native English speaking children. The children were selected because (a) the English school teacher approved of the project and allowed videotaping in the classroom; (b) parental permission was given; (c) the students were the correct age for their grade placement; and (d) they were characterized by the bilingual teacher as being Spanish dominant.

There were two bilingual teachers and one student teacher. The first one (Teacher A) taught the morning and afternoon classes for all but the last seven weeks of the school year, and because of sickness was replaced by Teacher B. Teacher

A was a native Anglo who had a "good" command of Spanish. She had been teaching in a bilingual program for three years. Teacher B, on the other hand, was Spanish dominant. She spoke English well, but with a marked accent. This was her first experience in a bilingual program, although she had been teaching for six years. Both teachers were assisted by the same student teacher who was of Mexican background. Both bilingual teachers stated that in all lessons, except English, only Spanish was used.

The two English teachers instructed their classes entirely in English. Maria's teacher stated that she encouraged the use of English in the classroom. She reported that the children only spoke English. Rosa's teacher on the other hand said that she encouraged the use of two languages in the classroom, and that she felt that both in and out of the classroom the Spanish dominant children spoke Spanish among themselves.

Videotaped data were collected three times during the 1975-76 school year--November 1975, February 1976, and May 1976. For each data collection period, each of the two target subjects were taped for an entire school day (half-day in the bilingual class and half-day in the English class).

In each taping session, the target child wore a wireless microphone. Two video cameras were used--stationary and portable. The stationary camera and a microphone were focused on the entire classroom in order to capture the child's placement and activities in the context of the whole class. The portable camera focused directly on the target child and her immediate social surroundings. This camera recorded the target child's conversation through the wireless microphone.

Bilingual Teacher A was present for the first two tapings. Bilingual Teacher B was present for the third. After the first taping session, all the teachers were interviewed to find out the teachers' language background, attitude toward bilingual education, and description of classroom language use patterns.

Table 1 presents the "approximate" amount of tape time we obtained for each lesson in the bilingual classroom for each time period. The English classroom times are also presented. While we attempted to videotape complete lessons, time was lost changing tapes, and fixing equipment. Also, as can be seen for taping period III, only half of Rosa's bilingual day and half of Maria's English day were recorded. Otherwise, there was a good sampling of lessons, most of which were complete. Recess was only taped for Maria's bilingual taping period III in the other individual activity times, children were working together on the same task.

TABLE 1

Situations Taped: Approximate Number of Minutes

	Bilingual a.m.						*Regular p.m.*
Maria	Math	Recess and Individual	English	Spanish	Culture	Total	All Subjects
Nov.	30	10	25	15	11	91	90
Feb.	20	None	20	40	10	90	75
May	15	15	30	35	05	100	45
	Bilingual p.m.						*Regular a.m.*
Rosa	Math	Individual	English	Spanish	Culture	Total	All Subjects
Nov.	10	10	15	15	15	65	120
Feb.	25	10	10	25	10	80	120
May	Not Taped	Not Taped	25	30	Not Taped	50	105

Data Analysis

A coding system was developed to analyze the video data. The final version satisfied three criteria: (a) it could easily be taught to a naive coder; (b) it required as little inference on the part of the coder as possible; and (c) it required a minimal amount of tape indexing and transcription.

In the bilingual classroom, every *interaction* in the *language environment* of the target children was coded. In the English classroom, only *interactions* between the target child and her peers were coded. An *interaction* is defined as a series of conversational turns by two or more speakers around a common activity or topic which is temporally linked. *Language environment* refers to the language that occurs during an activity in which the target child is participating or is present. Thus, in a lesson where the target child is participating, all interactions were coded--this includes interactions between the teacher and another child. If the target child was participating in one activity and the microphone picked up a conversation between a teacher and child involved in another activity, this interaction was not coded. Each interaction was coded by (a) initiator, (b) addressee, (c) number of seconds of talk in Spanish, (d) number of seconds of talk in English, (e) context of the interaction (a statement about the lesson, the group, the topic), and (f) gloss of interaction (a brief description of the interaction). Every time code-switching occurred this was noted, and the entire interaction was transcribed.

A detailed description of this coding system was written, along with the codings for two 30-minute tapes. The system was then taught to a research assistant who coded the tapes for all three data collection periods. A second assistant transcribed all the instances of code-switching. The codings were checked by the investigators who then used this system as a means of answering the research questions.

FINDINGS

Originally, the study was designed to compare language use patterns of bilingually schooled children across three time periods to determine whether and in what ways these might change in the course of a school year. However, because the bilingual teacher who was present for taping periods I and II was replaced by a second teacher for taping period III, the above comparison was not possible. Differences that might have occurred between the last taping period and previous ones might have been due to teacher rather than to time differences. Specifically, teacher differences might have been attributable to differences in the teachers' language abilities: Teacher A was English dominant, while Teacher B was Spanish dominant. Thus, what originally appeared to be a disaster for our design (i.e. a teacher change) became an added benefit in that a new and important dimension was added to the study: *teacher language ability*. Therefore, two types of comparisons were made: (a) a time comparison (time I vs. time II) and (b) a teacher language dominance comparison (time I and time II vs. time III).

Quantitative as well as qualitative analyses were performed on the data. The quantitative analyses were performed to determine how much of each language was used, and the qualitative analyses were performed to determine under which conditions each language was used.

Quantitative Analysis

The amount of Spanish and English used in the bilingual classes during teacher-child interactions.

Table 2 on page 46 contains the percentage of time Spanish was used in the language environments of both students in the bilingual classroom. In terms of *total amount* of time (last column) devoted to Spanish, there was a sizable decrease in the percent of time Spanish was used in the language environment of Rosa, while there was no change in amount of Spanish used for teacher-child interactions for Maria. It is unclear why there was a change in the teacher-child interactions from

time I to time II for Rosa's class only. One possibility is sampling error. As can be seen from Table 2, for taping period I, the amount of Spanish used by the teacher with Rosa's class (71 percent) is highly discrepant with the amount she used with Maria's class (41 percent). The latter figure, however, is very consistent with the amount of Spanish used in time II for both Maria's class (41 percent) and Rosa's class (46 percent). Thus it is possible that for some reason, the teacher just used more Spanish for that one particular taping period, which was not reflective of her normal patterns, and therefore the change noted from time I to time II in Rosa's class could be due to sampling error, rather than to natural time factors.

By time II, more time was being spent in English than Spanish. It should be noted that by February the entire culture lesson was taught in English even though it was designated by the teacher as a Spanish subject. More English also occurred in the Spanish language art lessons, and math lessons during time II than in time I.

There are differences between the amount of Spanish used in Teacher A's classroom and Teacher B's classroom, where the effect is most observable for Maria's class. (The data were too sparse to analyze for Rosa's class.) Teacher B clearly restored a Spanish balance to the classroom in math, Spanish, and culture, where most of her interactions with the students were in Spanish. Even during English language arts, she used more Spanish than Teacher A.

- *The amount of Spanish used during peer-peer interactions in the bilingual classes.*

In Teacher A's classroom less Spanish was used by Maria for peer interactions during time II (35 percent) than time I (93 percent), while there was not much change in the total amount used by Rosa. (This slight increase may not be real in that peer-peer interaction was not measured for the same activities for time I and time II.)

This decrease in Spanish for Maria might be caused by a number of factors. One possibility is that the children were becoming more proficient in English and thus used it more than Spanish in time II. A second possibility is that at time I the children were not attuned to the rules of the language use in the bilingual classroom which involved using more English than Spanish. By time II, their language use was consistent with these norms.

When Teacher B took over the class, there was an increase in the amount of Spanish used by the children among themselves. This increase in the use of Spanish at time III over time II was still not sufficient to reach the level of Spanish used by the children during time I. (See Table 3 on page 46.)

- *A comparison between language use by the children in the bilingual classroom with language use by the children in the regular classroom.*

Table 4 presents the percentage of time that the target children were engaged in peer interactions in Spanish in the regular and bilingual classes. Both children tended to use Spanish more in the bilingual classroom. (N.B.: Rosa uses equal proportions during time II.) By the end of the year, both children used relatively little Spanish in the regular classroom. In the bilingual classrooms, there was also a decrease in the amount of Spanish used from time I to time III by Maria. Thus Spanish was being used less, and English was being used more as the primary language of communication between peers in both the Spanish and English classrooms.

To summarize, language ability of the teacher is an important factor in the quantitative analysis of language use patterns in the bilingual classroom. In this study, when the English dominant teacher was present, the language of the classroom was predominantly English, while the opposite was true for the Spanish dominant teacher. The teacher's language ability also seems to affect the amount of Spanish used among the children in the bilingual although not the regular classroom.

Some of the data also suggest that time is an important factor in terms of amount of Spanish being used. Less Spanish was used for teacher-child interactions during taping period II in Rosa's classroom. Particularly interesting is the almost nonuse of Spanish during the time II culture lessons in both classrooms.

TABLE 2

Percent of Time Spanish Was Used for Teacher-Child Interactions in the Bilingual Class

Maria's Class	Math	Individual	English	Spanish	Culture	Total**
TIME 1 *Teacher A*	55	28	1	100	67	41
TIME 2 *Teacher A*	55	none on taping day	3	92	7	41
TIME 3 *Teacher B*	95	none on taping day*	29	97	99	71
Rosa's Class						
TIME 1 *Teacher A*	67	64	0	97	96	71
TIME 2 *Teacher A*	45	29	9	69	0	46
TIME 3 *Teacher B*	not taped	not taped	0	60	not taped	

**Recess taped in this period was not analyzed due to poor audio.*

***These totals are calculated by dividing the total time of Spanish by the total English and Spanish times.*

TABLE 3

Percent of Time Spanish Is Used During Peer-Peer Interactions in the Bilingual Class

Maria's Class	Math	Individual	English	Spanish	Culture	Total
TIME 1 *Teacher A*	100	25	no interactions	100	no interactions	93
TIME 2 *Teacher A*	0	no activity on taping day	89	25	100	35
TIME 3 *Teacher B*	13	not analyzed	70	35	96	50
Rosa's Class						
TIME 1 *Teacher A*	no interactions	56	no interactions	no interactions	no interactions	56
TIME 2 *Teacher A*	0	72	0	55	no interactions	67
TIME 3 *Teacher B*	not taped	not taped	33	90	not taped	

TABLE 4

Percentage of Spanish Used for Peer-Peer Interactions in the Bilingual and Regular Classes

Maria	Regular	Bilingual
Time 1	44	93
Time 2	1	35
Time 3	9	50

Rosa	Regular	Bilingual
Time 1	23	56
Time 2	66	67
Time 3	11	*

**It is misleading to give a total here as the data are based on only two activities, while the other taping periods are based on five activities. See Table 3 for breakdowns.*

Generally, less Spanish was used during peer-peer interactions in both the English and bilingual classes during time II than in time I.

Qualitative Analysis

While the above data reveal in gross terms how much each language is used in the classroom, they do not show how the languages were used nor the relative importance of the two languages for the participants in the classroom. A *qualitative* analysis was thus performed on the data to determine for what purposes and under what conditions each language was used in the classroom. This analysis involved the analysis of when and why code switching occurs. *Time* and *teacher* comparisons were again made.

●*Teacher-child interactions in the bilingual classroom.*

The qualitative analysis of teacher-child interactions was carried out along three dimensions: activity, person, and purpose of the message.

Activity. The use of two languages was examined during and between different activities. During period I, Teacher A used predominantly Spanish during Spanish language arts, math, and culture lessons only (see Table 2 on page 46.) English was used during English language arts.

English was also used by Teacher A at the very beginnings and ends of lessons during taping period I. (These time periods between lessons will be called transitions.) This was true even if the lesson was conducted in Spanish. For example, the last utterance of a math lesson was "Sit down and turn your papers over." The preceding and proceding utterances were all in Spanish. Of the 13 transitions that were taped for period I, eight were in English, three were in Spanish, and two were mixed (i.e. the two languages were used). The three Spanish utterances were made by the student teacher (who is Spanish dominant), while all eight English utterances were made by the bilingual teacher. Thus, during taping period I, the bilingual teacher was using English to mark junctures between segments of the school day. The use of English in these situations takes on the added meaning of a "contextualization cue" (cf. Gumperz, 1976), and implicitly serves the function of informing the participants that something new is about to happen.

In taping period II, the picture was not so clear. English was used not only in English language arts, but also for *culture*. Spanish predominated again for Spanish language arts. The situation with language use during transitions was also not so clear-cut as it was for the first taping. Of the seven transitions taped, four were in Spanish, two were in English, and one was mixed. It seems clear that language use patterns changed in the classroom. It may be that the teacher was no longer using English to mark junctures. This unclear pattern may also be a result of poor sampling procedures, since not enough transitions were taped from this period to arrive at any definite conclusions.

There were differences in the way Teacher B used language during and between activities from the way Teacher A used language. Teacher B used mainly Spanish for all lessons except for math and, of course, for English. Also, Teacher B did not use English to mark the transition from one activity to the next. Rather, the language used during transitions was usually the language of the preceding lesson, which in most cases was Spanish.

In summary, depending upon the taping period--and the teacher--specific languages were used for specific lessons. In addition, during taping period I English was the language most often used during transition periods. But, while "activity" is a useful dimension that accounted for some of the teacher-child language use patterns, it does not account for all of the variation. Other variables, namely "person" and "purpose," were examined, in an attempt to explain the language use patterns of this class.

Person. Language use in the bilingual classroom can be partially explained by consideration of the person with whom the teacher was interacting. Our analysis focused on the teacher's interaction with the student teacher, the two

target subjects, and other children in the class.

Teacher A only used English when speaking with the student teacher and vice-versa. However, Teacher B used Spanish with the student teacher. Again, the status and function that each language plays in the classroom is dependent on the teacher's language dominance.

Next the interactions between the teacher and each of the target subjects were analyzed. Table 5 (following) summarizes these data for Teacher A. In the case of Maria, there is a notable change in the language used from time I to time II. During time I, most of the teacher's initiated interactions with Maria were in Spanish (except during English, when only English was used); however, by time II, all interactions initiated by the teacher were in English. This change is also consistent with the change in Maria's initiations with the teacher. During time I, all initiations (except during English period) were in Spanish; by time II, all initiations were in English.

During time I, all of Rosa's interactions with the teacher were in English (except during Spanish). However, the teacher used mostly Spanish with her. It is puzzling that Rosa used so much English and yet was classified as Spanish dominant by the teacher. It could be that by switching to English, Rosa was more likely to get the teacher's attention and gain the floor without being sanctioned. By time II, all of the teacher's interactions with Rosa were in English. As there are only examples of Rosa's initiated interactions during Spanish language arts, it is impossible to make general statements about her language of initiation.

Thus, Teacher A began to use primarily English in interactions with the two target subjects by time II. This shift in Teacher A's and Maria's and Rosa's language of initiation probably reflects both the general Anglicization of the bilingual classroom and the teacher's recognition of these two children's increased ability to comprehend and communicate in their second language.

Teacher B was very consistent in her language use with the two target children (see Table 6 on page 50). She initiated in English for English language arts and in Spanish for all other activities. The target children were not so consistent. They had become used to the previous general use of English in the classroom, and unsuccessfully tried to interact with the teacher in English. She, however, pressured them to use Spanish by not answering their English questions (outside of English language arts) or by continuing to speak Spanish.

Sometimes, the teacher was unsuccessful in getting the child to use Spanish as in the following example:

Teacher: ¿Dónde está número dos?
Rosa: This is one, I did number five already.
Teacher: Ester, ésa es mala, mira, cambia ésta y ésta.
Rosa: Which one is wrong?

Teacher: Las dos están malas.
Rosa: Which one?
Teacher: Las dos. Allí están.

And sometimes she was successful:

Maria: I've done it already.
Teacher: Este no estaba bien. Está bien (etc.)
Maria: Esto no lo hice.

Occasionally Teacher B used English when a child initiated in English. However, she gave a minimal reply in English and then switched to Spanish:

Teacher: ¿Qué es tiza?
Rosa: Chalk.
Teacher: Chalk, right.
Lo que uno usa para escribir
en la pizarra.

Maria: Can I sharpen my pencil?
Teacher: No, that's sharp enough.
Y ahora vamos a leer la
página 64.
¿Comprende?

TABLE 5

Percentage of Interactions in Each Language Initiated by Teacher A and Target Children

Maria	MATH			ENGLISH			SPANISH			CULTURE			TOTAL		
Nov.	Eng	Sp	Cs*	Eng	Sp	Cs	Eng	Sp	Cs	Eng	Sp	Cs	Eng	Sp	Cs
Teacher Initiates with Maria	25	50	25	100	0	0	0	100	0	0	0	0	29	57	14
Maria Initiates with Teacher	0	0	0	75	0	25	0	100	0	0	100	0	50	33	17
Feb.															
Teacher Initiates	50	0	50	100	0	0	100	0	0	0	0	0	71	0	29
Maria initiates	100	0	0	67	0	33	0	0	0	100	0	0	83	0	17
Rosa															
Nov.															
Teacher Initiates with Rosa	33	67	0	100	0	0	0	100	0	0	0	0	18	82	0
Rosa Initiates with Teacher	100	0	0	100	0	0	0	100	0	100	0	0	53	47	0
Feb.															
Teacher Initiates	100	0	0	0	0	0	50	0	50	100	0	0	86	0	14
Rosa Initiates	0	0	0	0	0	0	0	33	67	0	0	0	0	33	67

**Cs=Code switching interactions*

TABLE 6

Percentage of Interactions in Each Language Initiated by Teacher B and Target Children

	MATH			ENGLISH			SPANISH			CULTURE			TOTAL		
	Eng	Sp	Cs*	Eng	Sp	Cs	Eng	Sp	Cs	Eng	Sp	Cs	Eng	Sp	Cs
Maria															
Teacher Initiates	0	100	0	100	0	0	0	100	0	0	0	0	50	50	0
Maria Initiates	20	40	40	100	0	0	67	33	0	0	0	0	62	23	05
Rosa															
Teacher Initiates				100	0	0	17	23	50						
Rosa Initiates				100	0	0	0	0	0						

**Cs=Code switching interactions*

TABLE 7

Percentage of Interactions in Each Language for Teacher-Child Interactions (Teacher A)

	MATH			INDIVIDUAL			ENGLISH			SPANISH			CULTURE			TOTAL TIME		
	Sp	Eng	Cs*	Sp	Eng	Cs	Sp	Eng	Cs	Sp	Eng	Cs	Sp	Eng	Cs	Sp	Eng	Cs
Maria's Class																		
Time 1	43	35	22	-	-	100	-	92	08	100	-	-	37	32	32	37	40	24
Time 2	11	55	33	none on taping day			-	64	36	20	80	-	-	67	33	03	59	43
Rosa's Class																		
Time 1	33	33	33	-	-	100	-	100	-	87	13	-	43	29	29	56	33	11
Time 2	13	63	25	-	33	67	-	-	100	18	18	64	-	100	-	12	45	42

**Cs=Code switching interactions*

TABLE 8

Percentage of Time Spanish Was Used for Interactions for Taping Period 2

	MATH	ENGLISH	SPANISH	CULTURE
Maria Period 2				
Teacher/Child	55	3	92	7
Child/Child	0	90	21	100

Thus, Teacher B attempted to use mainly Spanish with the two subjects.

By taping period II, it became clear that Teacher A used Spanish with specific students in the classroom, regardless of the activity going on. Those children she used Spanish with did not speak English very well. Clearly the two target children --Rosa and Maria--were not among this population. This suggests that Teacher A was becoming more flexible in her language use in order to meet the communication needs of individual children. This tendency is also seen with Teacher B, who during English language arts, used Spanish exclusively with a few children.

Thus, it is possible that the language chosen by the teacher was determined in part by her concept of the child's language ability. For Teacher A, this involved the use of English with those who understood it, and the use of Spanish for those who did not. For Teacher B, the rule was different: use Spanish with all children during Spanish, culture, and math; use English during English language arts, except use Spanish for those who can't understand English.

Purpose of the Message. While some language use patterns may be explained by "activity" and "person," a great many instances, especially those where code switching occurred, still remain unaccounted for. Many of these can be explained by focusing on the *content* or *function* of the utterance.

For time I, the teacher used English to comment on what was going on in the classroom, for giving directives, and commenting on the appropriateness or correctness of the students' work. Spanish was used in all lessons (except English language arts) by the teacher for providing information, facts, opinions, ideas, etc. While English was sometimes used for this purpose these cases occurred only when the informing in English was preceded by a directive in English. Abundant examples of this differential language use were found in the math lessons:

> O.K. let's go to the next one. ¿Cuánto en total?
> Siete, escribe "siete" en el cuadrito.
>
> Sit down! Turn your paper on the other side.

During time II, the explanation presented above becomes a little bit more complex. As can be seen from Table 7 on page 50, the proportion of teacher-child interactions where code switching occurred increased considerably from time I (Maria: 24 percent to 43 percent; Rosa: 11 percent to 42 percent). This jump is probably due to the teacher's more flexible use of the two languages in the classroom: she was more sensitive to the linguistic needs of individual children. Evidence for this is seen in the increase of code switching during the English lesson. Many of these code switches are explained by the teacher's translating words or directions from English into Spanish (or the other way around).

> And who remembers what this is? *Trigo*--wheat.
> When you're ready put up your hand and I'll check your papers.
> *Levante la mano.*

Usually translation occurred in interactions with individual children rather than with the group. During time II, English was still used for directives and procedural instructions or questions (e.g., "Can you see?") and Spanish for informing, especially in the math lesson. But again, the teacher seemed to modify this pattern by using the language that most suited the child's linguistic skills. Teacher B used some of these same patterns; however, the effects were not as clear since she did not use very much English. However, when she did use English outside of the English lesson, it was to give directives or to make procedural statements. Teacher B also used a lot of Spanish in the English lesson to translate instructions and concepts she had just given in English into Spanish.

> The eyes of the clown, please.
> *Solamente los ojos.*

These remarks were directed to individuals rather than to the class as a whole.

Thus, to some degree, the two teachers used Spanish and English to express similar functions, although not to the same extent. Teacher B seemed to be more consistent in her use of the two languages to express different functions than Teacher A. Teacher B's main rule for language use seemed to be: use Spanish as much as possible (except in the instances noted above). For Teacher A, a combination of the rules based on activity, person, and content discussed earlier seem to provide an adequate description of her language use strategies.

•*Qualitative analysis of peer-peer talk in the English classroom.*

In the regular classroom, the rule for differential language use between the target subjects and their peers seemed to be: use Spanish when the addressee's knowledge of English is extremely limited, otherwise use English. For all three taping sessions, Maria interacted in Spanish exclusively with one child, Miriam. In fact, the teacher used Maria during the first taping session as Miriam's interpreter. Her decrease in Spanish use in the regular classroom is a function of a decrease in her interactions with Miriam (see Table 4 on page 46).

During taping period I, Rosa interacted with a Spanish dominant child, Patty, solely in Spanish. She also served as her interpreter for the teacher. For the second period, Rosa was seated next to a new Spanish speaking child, Omar, with whom she only spoke Spanish. During this period, she also spoke exclusively in Spanish to Elena. Then in taping period III, a strange phenomenon occurred: Rosa was seated next to another child, Juan (who was also in the bilingual class). She began all her interactions with him in Spanish, but he replied in English. Eventually, by the end of the lesson, she only interacted with him in English. Clearly, Rosa was not following the standard rule: only use Spanish if the addressee cannot function in English. Why did she persist to speak to Juan in Spanish? There are several possibilities.

(a) Throughout the year in the English classroom, Rosa sat next to Spanish dominant children with whom she had interacted exclusively in Spanish. The situation with Juan marked the first time in our tapes that this pattern did not hold true. She may have spoken Spanish to Juan because she was used to being seated next to Spanish dominant youngsters with limited command of English. Unfortunately, we do not have sufficient information to confirm this hypothesis. (How long had she been sitting next to him, and why had the teacher placed her there?)

(b) There may have been some effect from the new bilingual teacher who had affected the language patterns of that classroom. The new teacher used Spanish as often as possible. Rosa may have learned that one should speak Spanish with the participants of the bilingual classroom regardless of the context and thus she spoke Spanish to Juan (who shunned these advances).

(c) It may also be that the new bilingual teacher caused some confusion in Rosa's sociolinguistic conception of the world--language use rules changed radically when the new teacher arrived. Rosa was having to learn a new set of rules for language use which still may have been in transition. There may have been mild confusion about which language to use in different situations, which was reflected in her inappropriate choice of language with Juan.

While there are many other interesting results, it was noted in our analysis of the regular classroom data that the bilingually schooled children were not marked for attending a special program. No negative remarks were made by either of the teachers or the other children in the classroom. Furthermore, there seemed to be a very lenient atmosphere in terms of language use. The children were never pressured to use English. In fact, the opposite sometimes occurred when the teachers used both students as translators for Spanish speaking peers.

•*Qualitative analysis of peer-peer talk in the bilingual classroom.*

The rules for peer-peer language use in the bilingual classroom were different from those in the regular classroom. These rules are not yet fully apparent to us since they appear to change among taping periods and differ for the two target children.

For time-period I, English was used by both Rosa and Maria while engaging in routinized and structured activities--reading a book, repeating well used phrases.

Spanish was used for negotiating an activity--choosing a book or setting up the rules of the game. This use of language by the students among themselves was in direct contrast to teacher language use in the bilingual classroom. The teacher used English to set up and negotiate situations, the students used Spanish; the teacher used Spanish for substantive parts of an activity, the students used English.

This reverse effect may be due to several factors. First, it may reflect the students' level of English language competence. It was easier to use English for routinized activities in that many of the utterances were set by the activity and had been used and rehearsed many times before. On the other hand, the use of Spanish was reserved for expression of more complex and more varied concepts. The second possibility was that the students used a pattern different from the teacher's to set themselves apart from her. This last explanation is given some credence by the data from period II, where the above patterns of English for routines and Spanish for negotiation still held for Rosa's data. On the other hand, Maria and her peers seemed to choose the opposite language from that used by the teacher in the lesson. Referring to Table 8 on page 50, we see this most clearly in the English and culture lessons, and somewhat in the Spanish lessons. This pattern is not found for Rosa's data.

For taping period III, the data do not seem to fit any of the above patterns. As mentioned earlier, language use patterns during period III were in the process of changing. The Spanish-negotiation, English-routine analysis of these data showed no systematic pattern. Using the opposite language of the teacher accounted for part of Maria's language use in math, English and Spanish, and for Rosa's in Spanish. However, many utterances still remain unexplained. We looked to see if the children used English or Spanish with certain children, but this did not yield positive results; we looked at the language that preceded the interaction and this, too, showed no regular patterns. In order to properly do this analysis, more detailed information on the children with whom the subjects were interacting would be required. However, this analysis indicated a number of interesting facts about peer-peer language use by bilingually schooled children: (a) there were a different set of rules for language use in the bilingual and regular classrooms; (b) in the bilingual classroom, these rules changed over the course of the year; and (c) it is possible that teacher's language dominance, which affected teacher-child interactions in the bilingual classroom, also affected child-child interactions in both classrooms.

DISCUSSION AND CONCLUSIONS

This study has described certain phenomena about the processes of bilingual schooling which should be of interest to both the educator and theoretician.

Several findings are of specific interest to the bilingual educator. The first is that language dominance of the teacher seemed to determine the language environment of the classroom. The English dominant teacher gave the bilingual classroom a distinctly English atmosphere, while the Spanish dominant teacher gave the classroom a Spanish atmosphere. Language dominance of the teacher affected language use patterns not only in terms of total amount of language used, but also in the ways each language was used. There was also the slight suggestion that teacher's language dominance might affect the child's language *outside* of the bilingual classroom (i.e. in the regular classroom).

Data such as these can be used by the bilingual educator to verify whether or not the type of program as set out by local guidelines is actually being followed by the teacher. In the context of the present study, Teacher A's language use patterns are consistent with the general guidelines for a "transitional" program. This is seen by her gradual increase of English into the curriculum over the year;

her speaking Spanish (during time II) only with children who could not function in English. Whether her behavior is intentional or not, these patterns are consistent with the philosophy of using the first language as a means (or bridge) for increasing facility in the second language, and then gradually phasing out the first language.

Similarly, these data are also of interest to those concerned with linguistic and cultural maintenance. One of the goals of many bilingual programs is to foster and promote growth in these areas. This was not achieved in this program. The children used less Spanish among themselves at the end of the year than at the beginning; the culture lesson (in Teacher A's classroom) was eventually taught in English rather than Spanish. However, these trends are fully consistent with the aims of a transitional program and offer further evidence that such a model was being followed.

These data have ramifications for teacher training programs. Because so many bilingual teachers are English dominant (General Accounting Office, 1976), they should be trained to become *more* conscious of their language use patterns so that they do not use English to the exclusion of the other language.

Lastly, this study is a good example of how process data can be crucial in the interpretation of product data. Specifically, in program evaluation, teacher language proficiency may be an important factor in accounting for pupil performance on certain tests. Children who have English dominant teachers (who, according to these pilot results, tend to use more English and less Spanish in the bilingual classroom) may perform better than expected on English language tests, and worse than expected on Spanish language tests. These results should not be attributed to failure of "bilingual education" to promote first language skills, but to individual teacher and classroom variation.

For the *sociolinguist*, many of the above results are also of interest for they shed some more light on the rules for speaking in different classrooms. The analysis indicated that not one, but a complex interaction of many social variables (e.g., activity, addressee, classroom, teacher language) must be considered in the discussion of language use patterns.

While certain primary patterns of language use were identified which confirm findings of other studies (cf. Shultz, 1975; Genishi, 1976; Phillips, 1975; Legarreta, 1977), many questions remain unanswered. The largest problem which deserves greater study is the patterns of peer-peer interaction. These are of importance because their interpretation will add insight into the social meaning these languages have for bilingually schooled children.

Results of this study also confirm other sociolinguists' (Blom and Gumperz, 1972) contention that self-reports of language use are highly unreliable. Specifically we have the case of Teacher A who reported that she used mainly Spanish in the bilingual classroom. Her perception of her language use was certainly not consistent with the data.

Lastly, we would like to consider the value of ethnographic and video data analysis as an evaluation technique. There is no doubt that this technique yields large quantities of rich data which can be analyzed in a number of different ways.* For example, in terms of the data collected for this analysis, the same tapes could also be used for (a) teacher training (pre- and in-service), (b) the assessment of academic skills in a naturalistic environment (e.g., reading), and/or (c) linguistic analysis of first and second language development.

However, for the *evaluator*, there are many difficulties in using videotape methods for collecting process data. What follows is a list of some problems which occurred in this study.

**Data on Rosa have already been used in a subsequent study investigating the relationship between personal, social, and cognitive utterances (as defined by the authors) and other variables (type of classroom, activity, person, language) in an effort to better understand the notion of self-concept (Walcer and Rodríguez-Brown, 1978).*

The first consideration is the *expense* in terms of both money and time. The equipment and the tapes are expensive. One must have a minimum of three staff in the classroom during taping (one for each camera, and a third taking notes on classroom events). In terms of time this study was very expensive: there were at least 18 full days of data collections (six days x three research assistants), another 21 days for coding, 10 days for transcription, and 17 days for data analysis. Thus, 66 working days were devoted to the study of two bilingually schooled students.

Despite the amount of data collected and the amount of man hours invested, we feel that our data are insufficient for making general statements about the nature of language use patterns in these classrooms. At several points in the paper we question the reliability of our data: Were the patterns seen for a particular taping period peculiar to that taping day only? Also, at times, results are not clearly interpretable because we lack important data regarding specific classroom variables (e.g., seating patterns, attendance, language abilities of nontarget subjects). Systematic on-site observations of the classrooms throughout the course of the year could have provided the necessary supplementary information. These problems indicate that video techniques are not sufficient to collect and analyze process data and that additional data collection procedures (and time) are required. Although the above procedures may be normal for the researcher, for the evaluator they are impractical. Those priorities involve collecting as much data as possible in as short a period of time as possible on the maximum number of subjects.

While it is true that a great deal of time and money were spent examining the school life of *only* two children, it is also true that the data collected are invaluable as they give us a lucid and detailed picture of the process of education for these students. Some compromise could be reached, so that the evaluator has access to these techniques. A system might be developed which cuts down on the time investment without greatly affecting the quality of the collected data. This system might include less use of videotape with use of audiotape and some on-site observations. Of course, the development of such a system could only result from a research project which had this development as its main aim.

REFERENCES

Austin Independent Schools. Bilingual Program Evaluation Report, ESEA, Title VII 1974-75. Austin, TX: Author, 1975.

Barik, H.C. and M. Swain. "Three-year Evaluation of a Large Scale Early Grade French Immersion Program: The Ottawa Study." Language Learning 25:1-30, 1975.

Beacon Independent School District. Bilingual Program Evaluation Report, ESEA, Title VII 1972-73. Beacon, NY: Author, 1973.

Blom, J.P. and J. Gumperz. "Social Meaning in Linguistic Structures: Code Switching in Norway." In Directions in Sociolinguistics, edited by J. Gumperz and D. Hymes. New York: Holt, Rinehart and Winston, 1972.

Bruck, M. and others. "Bilingual Schooling through the Elementary Grades: The St. Lambert Project at Grade Seven." Language Learning 24:183-204, 1974.

Chula Vista. Bilingual Program Evaluation Report, ESEA, Title VII 1972-73. Chula, Vista, CA: Author, 1973.

Cohen, A.D. A Sociolinguistic Approach to Bilingual Education: Experiments in the American Southwest. Rowley, MA: Newbury House, 1975.

Corpus Christi. Bilingual Program Evaluation Report, ESEA, Title VII 1974-75. Corpus Christi, TX: Author, 1975.

Cox, L. Caribou Bilingual Program Evaluation Report, ESEA, Title VII 1974-75. Caribou, ME: Caribou School District, 1975.

Dillon, P. "Annual Title VII Evaluation for San Francisco." 1971.

Erikson, F. and J. Shultz. "When Is a Context? Some Issues and Methods in the Analysis of Social Competence." Unpublished ms, Harvard University, 1976.

Genishi, C. "Rules for Code-Switching in Young Spanish-English Speakers." PhD dissertation [unpublished], University of California, Berkeley, 1976.

Goodenough, W. "Cultural Anthropology and Linguistics." In Language in Culture and Society, edited by D. Hymes. New York: Harper and Row, 1964.

Goodman, F.M. and C. Stern. Bilingual Program Evaluation Report, ESEA, Title VII 1970-71. Compton, CA: Compton City Schools, 1971.

Gumperz, J. "Language, Communication and Public Negotiation." In Anthropology and the Public Interest, edited by P.R. Sunday. New York: Academic Press, 1976.

Hymes, D. "Ethnographic Monitoring." Paper prepared for the Symposium on Language Development in a Bilingual Setting, Pomona, CA, 1976.

Legarreta, D. "Language Choice in Bilingual Classrooms." TESOL Quarterly 11(1): 9-16, 1977.

McDermott, R.P. "Kids Make Sense: An Ethnographic Account of the Interactional Management of Success and Failure in One First-Grade Classroom." PhD Dissertation [unpublished], Stanford University, 1976.

Mehan, H. and others. "The Social Organization of Classroom Lessons." A technical report submitted to Ford Foundation, 1976.

Phillips, S. "Code Switching in Bilingual Classrooms." MA Thesis [unpublished], California State University, Northridge, 1975.

Shultz, J. "Language Use in Bilingual Classrooms." Unpublished ms, Harvard Graduate School of Education, 1975.

_____. "It's not Whether You Win or Lose, It's How You Play the Game: A Microethnographic Analysis of Game Playing in a Kindergarten/First Grade Classroom." Working paper, Newton Classroom Interaction Project, Harvard Graduate School of Education, 1976.

U.S. General Accounting Office. Bilingual Education: An Unmet Need. Washington, DC: USGPO, 1976.

Walcer, C. and F.V. Rodriguez-Brown. "Social and Psychological Aspects of Language Use by Bilingual Children: An Ecological Analysis." Arlington Heights, IL: Bilingual Education Service Center, 1978.

Assessing Bilingual Programs: Methodological Considerations in Large-Scale Evaluations

Andrew D. Cohen
Flora V. Rodríguez-Brown

There are probably a number of bilingual programs across the U.S. and elsewhere for which extensive data have been collected but only partially analyzed, if at all. Excellent models for longitudinal evaluation such as that by Lambert and Tucker (1972) have been available for some years, but with some exceptions (e.g., Cohen, 1975; Cohen, Fathman, and Merino, 1975), have not generally been followed. Our purpose is not to advocate a pick-up-the-pieces or *post facto* approach to evaluation--but rather to provide one example of what can be done to obtain some research findings from seemingly disparate data as we did in downstate Illinois.

The downstate Illinois experience in assessment of bilingual programs in many ways reflects that of all U.S. school districts. Although there were initial talks of a design for a comprehensive longitudinal evaluation, this design was not carried out as planned. The data on downstate Illinois bilingual programs from their inception in 1971 can best be described as patchwork--some scales are available for some children at some points in their education. From the evaluator's perspective there appear to be mainly gaps. There are yearly statements of what evaluation *was* to be conducted, but little accountability for the actual testing (Seelye and Balasubramonian, 1973, being one exception). Furthermore, the amount of data collected far exceeded the amount or adequacy of data analysis performed.

The first step in the Illinois downstate evaluation was to locate and describe all existing bilingual evaluation data from Fall 1971 through Spring 1975. As it turned out, students' records for the first year, 1971-72, were not complete enough with respect to basic information to allow their inclusion in the matching program aimed at identifying participating students across school years. Consequently, the effort actually began with 1972-73 data. Although there were still missing data across students and across skill areas, the need for insights into the effectiveness of state bilingual programs warranted the analysis of these data. It was felt that qualified answers would emerge from such analysis (i.e. an analysis based on scores for some children on some measures at some points in time).

The second step was to form a database from the usable data. Suffice it to say here that the effort was time consuming and expensive and could have been avoided had a database been established at the outset of evaluation (i.e. Fall 1971). Once those data were collected and put into the database, then the decision was made as to what issues were accessible for discussion, given the data. Clearly, there are many questions one would like to ask about bilingual programs, but in the case of retrospective data, limitations on the data restricted the range of questions for which answers could be obtained.

It is, unfortunately, rather common for program evaluation to be a "pick-up-the-pieces" effort, taking place after the program is well into the implementation phase. This situation is sometimes (as in Illinois) provoked by the inconclusive efforts of earlier evaluators. Furthermore, it may well be the case that the bilingual program does not have clearly defined characteristics to evaluate. Thus, the results of even a well designed evaluation may be difficult to interpret.

Thanks to Taweewat Pitayanon for his assistance in statistical analysis.

When working at the cross-district level, as in downstate Illinois (15 selected districts), rather than at the level of one school district or even one individual school (unit) within that district, there is the further reality that the bilingual treatment is an assortment of treatments, sometimes changing in nature several times during the school year. Part of the task at hand was to attain consensus as to the principle characteristics of downstate Illinois bilingual programs.

In part, because of limitations on existing data and in part, out of a desire to continue the research effort longitudinally, new data were collected during the 1975-76 school year, both from subjects already having records in the database and from new subjects. By Spring 1976, the database provided opportunities to assess language ability, achievement, self-concept, and the relation between home language use and student language performance at school.

The retrospective data consisted of product (the outcomes of bilingual schooling) rather than process measures (assessment of the means of achieving outcomes). While most of the product data were collected at the end of the school year, some were collected at several intervals over the course of the year (Fall, Winter, and Spring). In brief, the aims of this study were: (a) to locate and describe disparate data from former years; (b) to add new data of specific interest to the best of these data; (c) to isolate a usable sample for continuing research; and (d) to make product statements, however qualified, about the effects of bilingual schooling on language ability, achievement, and self-concept. Generally, we attempted to evaluate with the intent of improving, not defending, ongoing programs in bilingual education in downstate Illinois.

Following are a series of specific research questions intended to give the presentation of findings a clearer, more precise focus. These questions are meant to reflect some general questions asked about bilingual education by a variety of different interest groups.

- *Do students who have participated in a bilingual program for several years achieve fluency and literacy in two languages?...What is the effect of number of years in bilingual schooling on English and Spanish reading achievement?...What is the effect of number of years in bilingual schooling on listening, speaking, and writing English and Spanish?* Questions of language proficiency concern legislators, teachers, and educational administrators. More specifically, it was the hope of educators in Illinois that bilingual programs would strengthen both languages, especially English. Whereas legislators in Illinois put emphasis on rapid and successful transition to English, they were also concerned about the effects of such programs on the maintenance of fluency and literacy in the student's home language.
- *Do students in a bilingual program achieve at a rate commensurate with their age, ability, and grade level in all subject areas?...What effect does years of bilingual schooling have on the development of general abilities--basic language concepts (word relationships, classifications, analogies) and mathematical concepts (computation and number series)?...What effect does years of bilingual schooling have on achievement in the content subjects (i.e. math, science, and social studies)?* Unfortunately, the primary question could not be answered directly with Illinois downstate data since there were no state norms for most tests (especially tests in Spanish), and many tests were not administered state-wide.
- *Do increased years in a bilingual program foster positive attitudes toward self, school, and community?* A basic tenet of bilingual schooling is that bilingual pupils' use of mother tongue, particularly in a program that enhances ethnic background, will concurrently instill within or reinforce positive attitudes toward self, school, and community. This study provided an opportunity to ask this question of a rather substantial cross-section of students.
- *How does the sociolinguistic environment at home affect language performance at school?...How does exposure to media, radio and television, in Spanish and English, influence language performance?...How does students' choice of language in speaking to mother, father, and siblings relate to language proficiency at school?* This is

just one set of questions attempting to relate background characteristics to performance outcomes. While the database actually offered an opportunity to relate many other background variables to performance outcomes, the present analysis was selected because it related sociolinguistic environment variables to performance outcomes in a rigorous way (see Fishman's [1977] discussion of the lack of research efforts to determine the direct influence of community and parental factors on achievement in bilingual programs).

THE CASE OF DOWNSTATE ILLINOIS

Subjects

Spanish-English bilingual pupils were selected for intensive study because they comprised the overwhelming majority of bilingual students downstate. The population of students entering into this evaluation consisted of two basic groups. The first was a cross-section of students in grades kindergarten through nine, for whom at least one piece of data was collected at any point between Fall 1972 and Spring 1975. The second group was a special group of 218 third and fourth graders who were identified as having been in a bilingual program for at least three years as of Fall 1975, and for whom longitudinal data already existed (group B-3, N=218) or as having just entered a bilingual program, at the third or fourth grade level in Fall 1975 (B-0 group, N=115).

Most of the information available on the population under study came from two questionnaires. One was administered to about 3,000 parents of students in grades kindergarten through three and to the students themselves in grades four and above, who were in bilingual programs during the 1974-75 school year. The second questionnarie was sent in Winter 1976 to parents of the 333 third and fourth graders tested during the 1975-76 school year (i.e. 218 students from the B-3 group and 115 students from the B-0 group) (see Cohen and Rodriguez-Brown, 1977, for copies of the questionnaires). Other information was obtained directly from the students' computerized records that accompanied test scores (i.e. information that would usually be filled out on the front of a test jacket). Such information included school district, grade and year of entry into a bilingual program, birth date, and sex. All of these data were entered into the database.

The pupils selected for longitudinal study attended bilingual education programs in 15 different school districts. The database provided us with a group of Spanish speaking third (N=109) and fourth graders (N=109) who were identified as having been in a bilingual program for at least three years as of Fall 1975 (referred to as the B-3 group). The rationale for choosing only these grade levels was to obtain a sample of children whose only schooling experience had been through bilingual education in downstate Illinois.

It was not possible to find a genuine control group since most, if not all, bilingual students were receiving some bilingual schooling during the 1975-76 year. The compromise approach was to use students who were just beginning bilingual schooling as a comparison group. Hence, a second group of third (N=66) and fourth graders (N=49) just beginning bilingual schooling was selected for comparative purposes (referred to as the B-0 group). In essence, the B-3 group formed a group for continuing longitudinal study and the B-0 group initially formed a comparison group; however, it was also the intent to longitudinally follow the B-0 group from their point of entry into bilingual programs.

As stated, in Winter 1976, the parents of the 333 children were mailed a bilingual questionnaire to fill out, 111 of the sets of parents of the B-3 group responded (51 percent), and 71 sets of parents of the B-0 group responded (62 percent).

It is important to indicate the major differences between the B-3 group (the group with three or more years of bilingual schooling at the start of the 1975-76 school year) and the B-0 group (the group just beginning bilingual schooling

at the third and fourth grade levels at that time). The very fact that the B-0 group was starting bilingual programs in grades three or four is an indication that these pupils included recent arrivals to those school districts, very possibly as immigrants from Mexico, Puerto Rico, or Cuba. Since this evaluation was intended as a continuing longitudinal study, a follow-up question would be, "Which initial differences between the B-3 and B-0 groups with respect to background characteristics disappear over time and which persist?"

As it turns out in this case, there were basic differences in baseline comparative data relating to the following areas: birthplace of the parents and children, parents' occupation, mother's schooling, and parents' and children's language skills and language use patterns. (Complete comparative data may be found in Ripley, 1976.)

Thus, the comparative picture is one of more established Spanish speaking families (the B-3 group) versus more recent immigrant families (the B-0 group), with the former having stronger English language skills, better educational backgrounds, and somewhat better jobs. Rather than avoiding comparison of B-3 and B-0 group students' performance at school because of these baseline differences, we felt that comparisons should still be made, but with careful attention to such initial differences as the longitudinal study progressed. The basic issue might actually concern the extent of the deficit that the more recent immigrants really have at the outset, given such background differences as those enumerated above. Just as bilingual schooling itself is an ever-changing process given the advances in the field, so are student and parental characteristics. These characteristics change with the assimilation and acculturation processes. Thus, the decision here was to consider these initial differences as important intervening variables to be monitored closely over time and to be considered seriously when interpreting any comparative results between the B-3 and B-0 group children.

During Winter 1976, questionnaires were also distributed to the teachers who taught the B-3 and B-0 group children included in the longitudinal study. The sampled teachers represented 15 school districts which had implemented bilingual education on or before September 1972. Thirty-two teachers out of 75 (43 percent) returned the questionnaires. These teachers had taught in their district's bilingual program on the average of four years.

The teachers indicated differences in the models of bilingual schooling that they employed in their classrooms. The most commonly used model was the half-day bilingual program, where children attended the bilingual program for half-a-day and the regular program for the other half; 39 percent of the teachers were involved in this approach. The second most common program was that of tutorial pull-out where the children attended the bilingual classroom for only an hour or so each day, sometimes receiving an English as a second language or a culture lesson (with emphasis on the culture of the ethnic group); 37 percent of the teachers indicated that they used this model. Other models reported were the team-teaching model (found in 13 percent of the classes), where the class had one bilingual and one regular teacher who taught together all the time, and the self-contained bilingual classroom (implemented by 11 percent of the classrooms), where the children spent all the time in a bilingual classroom with a bilingual teacher. Twenty-four teachers (75 percent of those responding) reported that they had a teacher aide in their classes as well, thus bringing the average student-staff ratio to 11 children per adult.

Three-quarters of the teachers reported that children who had been in the bilingual program since its inception received beginning instruction in English and Spanish reading concurrently.

Regarding language use in the classroom, 25 teachers (78 percent) specified that they used Spanish in their classroom primarily for Spanish language arts instruction, as a medium of instruction for all subjects, and for general classroom interaction. They reported speaking Spanish approximately 30 percent of the time on the average. With respect to student language use patterns at school, they reported that on the average 52 percent of their students spoke English and Spanish about half the time,

29 percent reportedly spoke Spanish more than English, and 19 percent reportedly spoke English more than Spanish.

Looking at the teachers' appraisals of their own language proficiency, 87 percent rated their spoken English ability as either "good" or "native," while 65 percent rated their spoken English ability as "native." In contrast, only 50 percent of the teacher aides were rated by teachers as having native command of Spanish, and 56 percent as having native command of English. Thus, the teachers appear to have been Spanish dominant on the whole, while aides were characterized by dominance in both directions. This finding regarding language dominance is in contrast to that of teachers and aides in other bilingual programs in Chicago and in the rest of the country especially in the Southwest (see Cohen, 1975). In parts of the Southwest, for example, parents were upset, at least at the outset of the bilingual programs, that most, if not all of the Spanish language instruction seemed to be provided exclusively by teacher aides, who were not considered to be as qualified as the teachers. Illinois is to be acknowledged for having recruited teachers with strong Spanish skills to teach in bilingual programs.

Measures

Between Fall 1972 and Spring 1976, a series of instruments, including tests of language dominance, reading, achievement in content subjects, and attitudes were given to children attending the Illinois bilingual programs.

The Inter-American English and Spanish Reading Tests (1-A Reading and Lectura). The test of reading and its Spanish equivalent *Prueba de Lectura* were developed under the direction of the late Herschel T. Manuel and are published by Guidance Testing Associates (1967, 1973). According to the author, these tests not only measure reading, but form a basis "for estimating ability to do school work in other areas in which the ability to read is related to achievement (Havassy, 1972)." The tests were developed by educators from Puerto Rico, Mexico, and Texas. The publishers report that "the language of the tests was chosen to avoid local idioms, and instead to use 'standard' language that could be understood generally (Guidance Testing Associates, 1967)." It is also reported that the English and Spanish versions of the test were checked for similarity of difficulty at all levels, grades K-12. Levels 2-4 of the Spanish version (Forms CEs and DEs) and of the English version (Forms CE and DE) were used in this study.

Level 2 has three subtests: Level of Comprehension (40 items), Speed of Comprehension (30 items), and Vocabulary (40 items). At this level, children choose a picture which is suggested by a printed word, a phrase, a sentence, or a paragraph. Levels 3 and 4 have the same subtest format as Level 2, but substitute written for pictorial stimuli. The tests are timed and group administered. Reliability coefficients (Kuder-Richardson 20) were calculated for both the English and Spanish versions' Level 2, using groups of downstate Illinois third and fourth graders (Fall 1976), respectively. In all cases, total test reliability was .95 or higher.

There were problems in deciding specifically which levels of the test to administer to the bilingual children in the study. The grade level designations were based on two criteria: (a) that children be native speakers of the language of the test and (b) that the language of the test be used actively in the environment in which the children live. For the Spanish version of the test, the first criterion was satisfied, but not the second, if we can assume that all the children were native speakers of Spanish and that English predominated in the community. For the English tests, the second condition was satisfied, but not the first. Due to this discrepancy, the children in the bilingual programs were given both Spanish and English versions of the test one level *lower* than the level corresponding to the grade that they were attending.

The Short Test of Linguistic Skills (STLS). The Short Test of Linguistic Skills (STLS) was developed by the department of research and evaluation of the Chicago board of education to measure language dominance in children whose native

language is not English. According to the authors, the test attempts to be "culture fair" in its content selection and considers the children's first language as a point of reference in isolating problems that students might have with English (Chicago Board of Education, 1976). The test is available in 11 parallel forms, including English, Arabic, Chinese, Greek, Italian, Japanese, Korean, Philipino, Polish, Spanish, and Vietnamese. The English and Spanish forms were used in this study.

The test is intended for administration to children in grades three through eight, and has four subtests: Listening, Reading, Writing, and Speaking. Each subtest has 20 items divided into four parts. Within each part, the items are ordered according to difficulty. The listening, reading, and writing subtests are group administered, and the speaking subtest is administered individually. The test was normed on 252 third graders and 248 fourth graders in Fall 1976. Kuder-Richardson 20 reliability coefficients ranged from a low of .83 (Spanish Reading, third graders) to a high of .97 (English Speaking, fourth graders).

The Illinois Inventory of Educational Progress (IIEP). The Illinois Inventory of Educational Progress (IIEP) was developed by the assessment and evaluation planning section of the Illinois office of education to provide the state with a comprehensive inventory of the state's educational progress in selected subject areas, at three key points in students' public schooling. The Inventory is primarily concerned with the assessment of "survival skills" in reading, mathematics, etc. According to the developers, "the IIEP provides for a systematic, continuous, census-like survey of knowledge, skills, and understanding established by students in three age groups (9, 13, and 17 years of age) and three grade levels (4, 8, and 11) (Illinois Office of Education, 1976)." The Inventory covers six different subject areas: Reading, Mathematics, Science, Social Studies, Writing, Career and Occupational Development, as well as some selected dimensions of the affective and psychomotor domain. This present study used only the Reading and Mathematics subtests.

Sequential Tests of Educational Progress (STEP). The Sequential Tests of Educational Progress (STEP) Series II is published by the Educational Testing Service. It consists of a battery of achievement tests designed to measure students' progress in academic areas. The STEP Listening and Writing Tests were selected for this study. According to the publishers, the tests are supposed to assess strength of performance, rather than speed, although they are timed. The publishers indicate the tests were standardized on a representative population within the United States, and norms are provided. Levels 2-4 of the test have been used in downstate Illinois, with Level 4 corresponding to grades 4-6, Level 3 to grades 7-9, and Level 2 to grades 10-12.

Boehm Test of Basic Concepts (Boehm). This test, published by the Psychological Corporation, is designed to assess children's knowledge of basic concepts. It is intended for administration to kindergarten and first grade students. The test was developed to identify and offer remedial help to children who do not have the linguistic-conceptual level of comprehension expected of them. The test has two parts, each having its own booklet (numbered one and two, respectively). There are both Spanish and English versions, forms A and B. The test consists of 50 sets of pictorial items which are organized in order of difficulty. For each item, the test administrator describes a depicted concept to the children and then instructs the children to mark the picture which corresponds to the concept. The test can be administered on a group basis and is not timed.

The Test of Basic Experiences (TOBE). The Language and Mathematics subtests from a battery of tests of basic experience published by McGraw-Hill were selected for administration. According to the authors, this test battery is designed according to the the premise that experiences and associated learning opportunities vary considerably among children. The test is based on the theory that for pupils to progress in school, certain concepts and skills which are often acquired before exposure to formal education must be mastered. The Level L form for kindergarten and grade one was selected for this study. There are both English and Spanish versions of the test. The publishers report an average Kuder-Richardson 20 reliability coefficient of .82 and assure content validity.

The Inter-American Tests of General Ability (I-A General Ability and Habilidad General). The tests of General Ability (Guidance Testing Associates, 1967, 1973) and its Spanish version *Prueba de Habilidad General* were also developed under the direction of the late Herschel T. Manuel and published by Guidance Testing Associates. According to the publishers, these tests are not intended to measure general intelligence, but are intended rather to provide an estimate of the ability to do academic work in general. The verbal subtest measures the understanding of written language and the ability to recognize relationships among concepts expressed by words. The nonverbal subtest tests for a grasp of relationships among concepts represented by pictorial stimuli. The mathematics subtest assesses the ability to think quantitatively through exercises in arithmetic computation and reasoning.

The test has six levels to it, each having an English and a Spanish version. This study used Levels 2-5. *Level 2* is to be administered to second and third grade children. The test consists of 100 items divided into two parts--a verbal numerical subtest (60 items) and a nonverbal subtest which consists of classification and analogies (40 items). *Level 3* is to be administered in grades four through six. It is a timed test which consists of 150 items. It has three subtests--verbal (sentence completion and word selection), nonverbal (figures, analogies, figure classification), and numerical (computation and number series). *Level 4* is for grades seven through nine, and consists of 150 items with the same format as Level 3. *Level 5* is for grades 10 through 13. It is timed, and comprises 150 items, with the same format as for Levels 3 and 4.

The Bilingual Test Battery (BTB). The Bilingual Test Battery (BTB) was developed by the department of research and evaluation of the Chicago board of education. The BTB is designed to assess both achievement in the content areas of math, science, and social studies, and attitudes toward self. Items were constructed on the basis of a review of both standardized and teacher-made tests and on the basis of program objectives as listed in school districts' proposals for the continuation of bilingual program funding.

The test has three levels: Level 1 covers grades three and four, Level 2 covers grades five and six, and Level 3 covers grades seven and eight. For each level, the 80-item test is divided into four subtests: Mathematics (20 items), Social Studies (20 items), Science (20 items), and Attitudes (20 items). The special feature of this test is that the items are presented to the children in Spanish and English concurrently, on opposite sides of the page. The children are encouraged to read the items in the language with which they are more familiar. The test is group administered and is not timed.

Self-Observation Scales (SOS). The Self-Observation Scales, developed by the National Testing Service, consist of a group administered instrument at Primary and Intermediate levels, with versions in English and in Spanish. The Primary Level (intended for grades kindergarten through three) consists of 45 items. It measures five dimensions of children's affective behavior: Self-Acceptance, Social Maturity, School Affiliation, Achievement Motivation, and Self Security. The Intermediate Level (intended for grades four through six) consists of 60 items. It measures the same five dimensions as on the Primary form and adds three more: Teacher Affiliation, Social Confidence, and Peer Affiliation.

The norms for the Primary Level were based on a sample of 9,030 students in grades kindergarten through three, and the norms for the Intermediate Level were based on a sample of 7,580 children in grades four through six. Special attention was paid to the social, geographic, and socioeconomic characteristics of the participating schools. In all, 150 schools nation-wide participated in the norming of the test.

As mentioned, this evaluation went through a variety of phases with little, if any, cohesive, continuing longitudinal overview. Thus, data were collected in what may seem like a disparate fashion. In reality, however, there were various short-term plans for data collection which were as much concerned with issues of test development (such as test reliability) as with the actual achievement

of the youngsters tested. There was a reluctance to report achievement scores before the reliability and validity of the instruments were established on Illinois bilingual students. A detailed description of which students received which measures when, is provided in Cohen and Rodriguez-Brown (1977).

The Formation of the Database

As of Fall 1975, there was an accumulation of data from students in bilingual programs downstate dating back to Fall 1971. Some data were on tape, some on cards, some simply in the original exam booklets. Most of the data were, in fact, out of state--in the hands of a small data analysis outfit in North Carolina (IBEX). The rest of the data were at the Bilingual Service Center in Illinois. Unfortunately, student identification numbers were purposely not preserved over the years when students were retested (if they were), out of a concern for confidentiality. Although a concern for anonymity is legitimate, there could have been other means of preserving anonymity without all but prohibiting longitudinal use of student records. Even in the face of the given reality, it was clear that if any statements were going to be made about the effects of bilingual programs over the years, there was a need to merge the existing data into a comprehensive database. This database could then serve not only as a repository from whence information could be available for evaluation and research, but also as a source of data for decision-making and proposal writing.

The Illinois Institute of Technology's (IIT) Research Institute developed routines for building such a database, using their DEC's PDP-11/45 (Ripley, 1976). IIT designed programs to build "logical records" for each student for whom there was data. The staff "built" records in the sense that they had to construct routines for searching through data from disparate testing sessions and years for multiple matches. Such matches, then, would indicate that the records belonged, in fact, to the same student. The outcome was a database with over 273 variables from 70 data files on a master tape, describing attributes and academic achievement of 4,698 students from grades K-12 in 44 school districts in downstate Illinois. The data span the four school years from 1972 through 1976. It was not until this elaborate matching procedure had been completed that we were able to identify just which third and fourth graders had three or more years of bilingual schooling as of Fall 1975. The database identified over 200 from 15 school districts. We then set out to locate the students, as well as to administer supplemental tests to them. These are the students referred to as the B-3 group.

Returning to the general format of the database, the data were thus set up on a personal student basis, with one identification number for each student. The database, then, had two sections, one with constant information such as school, district, birth year, year entered bilingual program, grade entered bilingual program, birth place of various family members, etc. The other section consisted of data that were added to over the years as additional data on the students were obtained. Such data included achievement scores and background data of a changing nature over the years--students' language ability and language use patterns.

Data Analysis

Data analysis procedures posed a problem in this study (as they have in many educational evaluation programs) since there was no carefully selected control group with random assignment of pupils to experimental and nonexperimental conditions. It was with this constraint on statistical analysis, and with the awareness of consequent limitations as to generalizability of results that statistical analysis was carried out.

A cross-sectional aggregate data approach to analysis seemed the most practical in that there was a large database with which to work. This cross-sectional approach, also called grade cohort analysis when working with school grades, calls for comparative analysis of groups of students at the same grade level who differ

on some characteristic, in this case on number of years of bilingual schooling. Data from single school programs usually are not ample enough to allow such analyses. In this instance, grade cohort analysis produced a compromise approach to the problem of no existing control group, allowing instead the possibility of comparing across students all within bilingual schooling programs, but with differing amounts of exposure time. This analysis procedure allowed for analysis not only of the B-3 and B-0 groups, but of other groups in the database as well.

The actual statistical techniques varied according to the extent of available data on the students concerned. One of the first analyses involved the computation of an intercorrelation matrix of test scores for all available data between 1972 and 1976. This matrix helped identify groups of individual students who had taken two or more tests at different times over the four years under investigation. The correlation matrix was also used to determine which tests clustered and which should therefore be used to measure which skills. (This correlation matrix appears in Ripley, 1976.) Other statistical techniques employed included T-tests, one-way analysis of variance with and without the Scheffé multiple range test, two-way analysis of variance, and two-way analysis of covariance, primarily using programs from the Statistical Package for the Social Sciences (Nie et al, 1975).

One-way and two-way analyses of variance were used for analysis. In cases where groups were assumed to be similar, analysis of covariance was used to adjust post-test score differences according to pretest score differences. Lord (1967) and subsequent researchers have pointed out that if groups are not random to begin with, strong evidence is needed to determine that the selection was random "in effect." It is difficult to say whether the downstate Illinois data met this assumption. It was nevertheless assumed that these pre-existing groups were sufficiently similar to warrant use of analysis of covariance. The SPSS two-way ANVOCA program did adjust for unequalcell sizes.

As noted above, it was documented that there were basic differences between background characteristics of the B-3 group (students with three or more years of bilingual schooling as of Fall 1975) and the B-0 group (students just entering bilingual programs at the third or fourth grade level in Fall 1975), selected for comparative analysis in 1975-76. In fact, these differences become part of the longitudinal analysis. The ongoing question was: If a group begins a bilingual program in, say, grade three or four because they are recent arrivals from Spanish speaking countries or elsewhere and have certain corresponding characteristics (more Spanish in the home, parents having somewhat lower-status occupations, education, etc.), what will the effect of these differences be on outcomes? Of particular interest was the *continuing* effect of these differences. Thus, this analysis in many ways simply established baseline data so that these students could continue to be compared with those having previous bilingual schooling.

The data analysis was subcontracted to the Illinois Institute of Technology's Research Institute (IITRI) and IITRI reported that because of time constraints and budget limitations, certain statistical analyses that would have further validated the statistical procedure utilized, were not carried out. Such analyses would have included checking for the parallelism of regression lines between groups in analysis of covariance (ANCOVA) and the use of multivariate analysis of variance procedures. Thus, no attempt was made to analyze related tests together, i.e. Spanish and English reading. The use of multivariate analysis would have indicated how much shared variance there is between presumably related tests.

DISCUSSION AND CONCLUSIONS

The findings from this study indicated that children in downstate Illinois bilingual classrooms were in certain ways better off for having received bilingual schooling (see Cohen and Rodriguez-Brown, 1977, for the details). For example, incremental years of bilingual schooling for minority group children enhanced their fluency and literacy skills in the dominant language of the society, English.

Students with more years of bilingual schooling were stronger in both the productive English skills (speaking and writing) and the receptive skills (listening and reading). Furthermore, incremental years of bilingual schooling appeared to enhance conceptual development in general. A sampling of different assessment measures all produced indications that bilingual schooling enhanced cognitive functioning on tasks assessing nonverbal conceptual skills and verbal concepts, particularly English language concepts. Incremental years of bilingual schooling also appeared to contribute to growth in specific subject matter areas as well. Students with more years of bilingual schooling performed stronger on math, science, and social studies achievement measures.

There were also findings less consistent with the aims of bilingual programs. First, a goal of bilingual programs in Illinois was to promote fluency and literacy in two languages; this study suggested that incremental years of bilingual schooling were not serving to enhance native language skill in all areas. Whereas the Spanish-English bilingual programs sampled appeared to be promoting Spanish listening and writing skills, speaking and reading skills did not appear to improve with increased years in a bilingual program (as assessed cross-sectionally). Second, another goal of the bilingual programs was to insure that every student would be proud of himself, his family, and his background. The findings in this study suggest that attitudes toward self, toward school, and toward community were negatively influenced by incremental years of bilingual schooling, at least at the elementary school level. It may be that the particular programs and methods that downstate Illinois programs have selected may be producing these negative consequences, at least among those students sampled.

The findings relating home language use to school language proficiency indicated the potential influence of native language use out of school on native language proficiency as measured at school. Students who did not speak Spanish to their mothers at home performed low in Spanish achievement at school. Those students who spoke Spanish with their siblings at home had high Spanish achievement scores. These findings simply underscore the role that out-of-school linguistic or sociolinguistic factors may play in school language achievement, regardless of the nature of the bilingual program. It may be that if societal forces are working against maintenance of Spanish, a subordinate language in the society, then even the best of bilingual programs will have difficulty promoting the maintenance of Spanish literacy and fluency skills among their students.

Again we must point out that whether the data met the assumptions behind the statistical techniques used in the analysis is debatable. Theoretically, irrefutable conclusions would have required a rigorous control group rather than a grade cohort approach, particularly concerning the B-3 and B-0 groups.

Other unknown variables also make interpretation of findings anything but definitive. First, there were no easy means of determining just what treatments the students received over their several years of bilingual schooling. There were no attempts to document approaches on a yearly basis, through specification of models, recording of classroom schedules, teacher reports of language use, etc. (see Cohen, 1975). When following large numbers of students, it may be that such documentation of treatments is simply impractical at the cross-district level. Another issue regarding these data results concerns the possibility that those students in bilingual programs who had spent more time in the program were more likely to have been given more tests, and had thus become "test-wise," possibly inflating their test scores slightly, particularly on repeated administrations of the same type of test. Of course, such children could also become tired of so much testing and thus not try as hard as students in a new control group.

There are other problems inherent in these analyses. Aside from the obvious consideration that achievement gains could be due to factors other than the treatment (i.e. bilingual schooling), there is also the confounding of school experience, age of student, and level of test. These analyses did not control for age which could explain some of the variance. Also, when controls were applied statistically

through ANCOVA, there was no check on the validity of the statistic through analysis of the parallelism of regression lines. Although thousands of dollars went into the data analysis phase alone (separate from the thousands spent on the formation of the database), a number of seemingly important analyses were not run. We were informed by the IIT Research Institute that they did not have time to mobilize complex programs (i.e. multivariate analysis) nor money to carry out such analyses. (One way to guard against incomplete analyses would be to supervise the analysis directly, rather than subcontracting it out as in this case.)

There is clearly a purpose for continuing longitudinal study of the B-3 and B-0 groups over subsequent years--if for no other reason than to see if the B-0 group catches up in English and in academic subjects, as well as to assess maintenance of Spanish skills. Also, the attitudinal data here suggest the importance of looking at attitudes over time rather than cross-sectionally by cohorts. Also, it would be worthwhile to relate attitudinal measures to achievement measures, again an analysis that was not included in the present study. Neither was multivariate analysis of variance run on groups of related tests, e.g., Spanish and English reading tests. Such analyses might reveal relationships that are masked by separately analyzing tests which are related, such as the Spanish and English versions of the Inter-American Reading Tests.

Particularly now that so much has gone into the formation of the database, time needs to be spent doing a variety of statistical analyses to provide even greater insights into the effects of bilingual schooling in downstate Illinois. There is also room for continued test development and the administration of new and different tests in the evaluation process. We also conducted a pilot study to determine Spanish reading progress through a criterion-referenced test developed professionally and marketed commercially--marking a compromise between a teacher-made test and a standardized test in terms of its proximity to the objectives of the local classroom.

It has been suggested that there should be more "tracer" studies to determine the probability that a pupil will go out into the community and function comfortably in English or bilingually (Tucker, 1977). Such a study would help determine whether students who leave bilingual programs are really functioning in English. Investigations could include types of higher education (if any), success at studies, jobs selected, skills required, proficiency at jobs, etc. Such follow up studies have been done linking graduate studies to job performance, but few, if any, have linked elementary school bilingual experiences to adult performance. But in order for such studies to be successful, there must be an efficient system of record keeping. The task is complicated by the sometimes high transiency rate among minority families who move around in search of work. Clearly, the model of a database as constructed for the present study could suggest to an evaluation staff a means of storing data over time.

In many ways this present study was an exercise in compromise, reflecting the trials and tribulations of evaluation in the nonlaboratory setting. Perhaps the biggest asset of the study was the creation of the database. This database now exists for use in future downstate evaluations and comparative evaluations within the state and across states. It is reassuring that this study is only one of several studies beginning to appear regarding the impact of bilingual schooling. But there is still much work to be done in the area of research design and data analysis (see Cohen and Laosa, 1976). And the ultimate comparability of research findings depends to a significant degree on the comparability of the research methods.

We have repeatedly drawn attention to areas in which this evaluation is vulnerable to criticism. The question really becomes one of the degree of compromise permissible in the effort to assess the impact of bilingual schooling. We feel that as long as methods and their shortcomings are made clear, then the reader can use the findings profitably. Hopefully, an accumulation of such reports over time will begin to paint an honest picture of the impact of bilingual schooling. For example, there may begin to emerge more definite conclusions as to the effects of such programs on English language acquisition, on native language maintenance, on attitudes toward self and community, etc. At the present time, the results are still "mixed."

REFERENCES

Chicago Board of Education. Short Tests of Linguistic Skills: Program Guide. Chicago: Department of Research and Evaluation, 1976.

Cohen, A.D. A Sociolinguistic Approach to Bilingual Education. Rowley, MA: Newbury House, 1975.

_____ and L.M. Laosa. "Second Language Instruction: Some Research Considerations." Journal of Curriculum Studies 8(2):149-165, 1976.

_____ and F.V. Rodriguez-Brown. "Evaluation in Moderate-to-Small School Districts: Downstate Illinois." Report to the Bilingual Education Service Center, Arlington Heights, IL, 1977.

_____ and others. "The Redwood City Bilingual Education Project, 1971-1974: Spanish and English Proficiency, Mathematics, and Language Use Over Time." Working Papers on Bilingualism 8:1-29, 1975.

Fishman, J.A. "The Social Science Perspective." In Bilingual Education: Current Perspectives, Vol. 1. Arlington, VA: Center for Applied Linguistics, 1977.

Guidance Testing Associates. Manual and Technical Report: Tests of General Ability and Tests of Reading. Inter-American Series. Austin, TX: Author, 1967.

_____. Supplement to 1967 Manual, Inter-American Series: California Report. Austin, TX: Author, 1973.

Havassy, B. "A Critical Review of the New Inter-American Series." In Proceedings of BABEL Workshop and Preliminary Findings, Multilingual Assessment Program. Berkeley, CA: Bay Area Bilingual Education League (BABEL), 1972.

Illinois Office of Education. The Illinois Inventory of Educational Progress: Teachers' Manual. Chicago: State Board of Education, 1976.

Lambert, W.E. and G.R. Tucker. The Bilingual Education of Children. Rowley, MA: Newbury House, 1972.

Lord, F.M. "Elementary Models for Measuring Change." In Problems in Measuring Change, edited by C.W. Harris. Madison, WI: University of Wisconsin Press, 1967.

Nie, N.H. and others. Statistical Package for the Social Sciences (second edition). New York: McGraw-Hill, 1975.

Ripley, T.C. Analysis of the Downstate Illinois Bilingual Education Database: A Final Report. Chicago: Illinois Institute of Technology Research Institute, 1976.

Seelye, H.N. and K. Balasubramonian. "Accountability in Educational Reform Programs through Instrumentation Analysis and Design Variation: Evaluating Cognitive Growth in Illinois Bilingual Programs, 1972-73." Arlington, VA: Available from EDRS, 1973 (ED 112 635).

Tucker, G.R. "The Linguistic Perspective." In Bilingual Education: Current Perspectives, Vol. 2. Arlington, VA: Center for Applied Linguistics, 1977.

U.S. General Accounting Office. Bilingual Education: An Unmet Need. Washington, DC: USGPO, 1976.

Selected Additions to the Literature on Bilingual Education

Alatis, J.E. and K. Twaddell, eds. English as a Second Language in Bilingual Education. Washington, DC: TESOL, 1976.

American Institutes for Research. Evaluation of the Impact of ESEA Title VII Spanish/English Education Programs, Volume I: Study Design and Interim Findings. Palo Alto, CA: Author, 1977.

Andersson, T. and M. Boyer. Bilingual Schooling in the United States (second edition). Austin, TX: National Educational Laboratory Publishers, 1977.

Berman, P. and M.W. McLaughlin. Federal Programs Supporting Educational Change, Vol. 1: A Model of Educational Change. Santa Monica, CA: The Rand Corporation, 1974.

Blanco, G. "The Education Perspective." In Bilingual Education: Current Perspectives, Vol. 4. Arlington, VA: Center for Applied Linguistics, 1977.

Bruck, M. and others. "Bilingual Schooling through the Elementary Grades: The St. Lambert Project at Grade Seven." Language Learning 24(2):183-204, 1974.

_____ and others. "Assessing Functional Bilingualism within a Bilingual Program: The St. Lambert Project at Grade Eight." Montreal: Department of Psychology, McGill University, 1975.

_____ and others. "Alternative Forms of Immersion for Second Language Teaching." Working Papers on Bilingualism 10:22-75, 1976.

Burt, M. and H. Dulay. "Some Guidelines for the Assessment of Oral Language Proficiency and Dominance." TESOL Quarterly 12(2):177-192, 1978.

_____ and other eds. Viewpoints on English as a Second Language. New York: Regents, 1977.

Chicago Board of Education. Chicago's Bilingual Programs Evaluation Report 1975-76. Chicago: Department of Research and Evaluation, 1977.

Cicourel, A.V. and other eds. Language Use and School Performance. New York: Academic Press, 1974.

Cohen, A.D. "Progress Report on the Culver City Spanish Immersion Program: The Third and Fourth Years." UCLA Workpapers in Teaching English as a Second Language 9:47-65, 1975.

_____. A Sociolinguistic Approach to Bilingual Education: Experiments in the American Southwest. Rowley, MA: Newbury House, 1975.

_____. "The Case for Partial or Total Immersion Education." In The Bilingual Child: Research and Analysis of Existing Educational Themes, edited by A. Simões, Jr. New York: Academic Press, 1976.

_____ and L.M. Laosa. "Second Language Instruction: Some Research Considerations." Journal of Curriculum Studies 8(2):1-17, 1976.

Cooper, R. "Research Methodology in Bilingual Education." Paper presented at the 29th Annual Georgetown Round Table, Washington, DC, March 16, 1978.

Cornejo, R.J. A Synthesis of Theories and Research on the Effects of Teaching in First and Second Language: Implications for Bilingual Education. Austin, TX: National Educational Laboratory Publishers, 1974.

Cummins, J. "The Influence of Bilingualism on Cognitive Growth: A Synthesis of Research Findings and Explanatory Hypotheses." Working Papers on Bilingualism 9: 1-43, 1976.

Duran, R., ed. Proceedings of the First National Conference on Chicano and Latino Discourse Behavior. Norwood, NJ: Ablex Publishing Co., forthcoming.

Epstein, N. Language, Ethnicity, and the Schools: Policy Alternatives for Bilingual-Bicultural Education. Washington, DC: George Washington University Institute for Educational Leadership, 1977.

Fishman, J.A. Bilingual Education: An International Sociological Perspective. Rowley, MA: Newbury House, 1976.

_____. "The Social Science Perspective." In Bilingual Education: Current Perspectives, Vol. 1. Arlington, VA: Center for Applied Linguistics, 1977.

_____ and others. The Spread of English. Rowley, MA: Newbury House, 1977.

Gaarder, A.B. Bilingual Schooling and the Survival of Spanish in the United States. Rowley, MA: Newbury House, 1977.

González, G. "Teaching Bilingual Children." In Bilingual Education: Current Perspectives, Vol. 2. Arlington, VA: Center for Applied Linguistics, 1977.

Hornby, P.A., ed. Bilingualism: Psychological, Sociological, and Educational Implications. New York: Academic Press, 1977.

Lambert, W.E. "The Effects of Bilingualism on the Individual: Cognitive and Sociocultural Consequences." In Bilingualism: Psychological, Sociological, and Educational Implications, edited by P.A. Hornby. New York: Academic Press, 1977.

Lewis, E.G., ed. "Bilingualism in Education--Cross-National Research." International Journal of the Sociology of Language 14:5-30, 1977.

Mackey, W.F. and T. Andersson, eds. Bilingualism in Early Childhood: Papers from a Conference on Child Language. Rowley, MA: Newbury House, 1977.

_____ and V.N. Beebe. Bilingual Schools for a Bicultural Community: Miami's Adaptation to the Cuban Refugees. Rowley, MA: Newbury House, 1977.

Macnamara, J. and others. "Attending a Primary School of the Other Language in Montreal." In The Bilingual Child: Research and Analysis of Existing Educational Themes, edited by A. Simões, Jr. New York: Academic Press, 1976.

McKay, M. "Spoken Spanish of Mexican American Children: A Monolingual and Bilingual School Program." PhD dissertation [unpublished], Stanford University, 1975.

Mehan, H. "Ethnography." In Bilingual Education: Current Perspectives, Vol. 1. Arlington, VA: Center for Applied Linguistics, 1977.

Parker, L.L. (and S.B. Heath). "Current Perspectives." In Bilingual Education: Current Perspectives, Vol. 5. Arlington, VA: Center for Applied Linguistics, 1978.

Paulston, C.B. "Ethnic Relations and Bilingual Education: Accounting for Contradictory Data." In Proceedings of the First Inter-American Conference on Bilingual Education, edited by R.C. Troike and N. Modiano. Arlington, VA: Center for Applied Linguistics, 1975.

_____. "Research." In Bilingual Education: Current Perspectives, Vol. 2. Arlington, VA: Center for Applied Linguistics, 1977.

Ramírez, M. and others. Spanish-English Bilingual Education in the U.S.: Current Issues, Resources, and Research Priorities (Papers in Applied Linguistics, Bilingual Education Series No. 5). Arlington, VA: Center for Applied Linguistics, 1977.

Rubin, J. "Toward Bilingual Education for Paraguay." Paper presented at the 29th Annual Georgetown Round Table, Washington, DC, March 16, 1978.

Saville-Troike, M. Foundations for Teaching English as a Second Language. Englewood Cliffs, NJ: Prentice-Hall, 1976.

_____. A Guide to Culture in the Classroom. Arlington, VA: National Clearinghouse for Bilingual Education, 1978.

Segalowitz, N. "Psychological Perspectives on Bilingual Education." In Frontiers of Bilingual Education, edited by B. Spolsky and R. Cooper. Rowley, MA: Newbury House, 1977.

Silverman, R.J. and others. Oral Language Tests for Bilingual Students: An Evaluation of Language Dominance and Proficiency Instruments. Portland, OR: Northwest Regional Educational Laboratory, 1976.

Simões, A., Jr., ed. The Bilingual Child: Research and Analysis of Existing Educational Themes. New York: Academic Press, 1976.

Skutnabb-Kangas, T. and P. Toukomaa. Teaching Migrant Children's Mother Tongue and Learning the Language of the Host Country in the Context of the Socio-Cultural Situation of the Migrant Family. Department of Sociology and Social Psychology, University of Tampere, Finland, 1976.

Spolsky, B. "Bilingualism." In A Survey of Applied Linguistics, edited by R. Wardhaugh and H.D. Brown. Ann Arbor, MI: University of Michigan Press, 1976.

_____ and R.L. Cooper, eds. Frontiers of Bilingual Education. Rowley, MA: Newbury House, 1977.

_____ and R.L. Cooper, eds. Case Studies in Bilingual Education. Rowley, MA: Newbury House, 1978.

Stern, H.H. and A. Weinrib. "Foreign Languages for Younger Children: Trends and Assessment." Language Teaching and Linguistics: Abstracts 10(1):5-25, 1977.

_____ and others. Three Approaches to Teaching French. Toronto: Ministry of Education, Ontario, and OISE, 1976.

Sumner, G. and G. Zellman. Federal Programs Supporting Educational Change, Vol. VI: Implementing and Sustaining Title VII Bilingual Projects. Santa Monica, CA: Rand Corporation, 1977.

Swain, M. and H.C. Barik. "Bilingual Education for the English Canadian: Recent Developments." In The Bilingual Child: Research and Analysis of Existing Educational Themes, edited by A. Simões, Jr. New York: Academic Press, 1976.

_____ and H.C. Barik. Five Years of Primary French Immersion. Toronto: OISE, 1976.

Troike, R.C. and N. Modiano, eds. Proceedings of the First Inter-American Conference on Bilingual Education. Arlington, VA: Center for Applied Linguistics, 1975.

Tucker, G.R. "The Linguistic Perspective." In Bilingual Education: Current Perspectives, Vol. 2. Arlington, VA: Center for Applied Linguistics, 1977.

_____ and G.A. Cziko. "The Role of Evaluation in Bilingual Education." Paper presented at the 29th Annual Georgetown Round Table, Washington, DC, March 16, 1978.

U.S. Commission on Civil Rights. A Better Chance to Learn: Bilingual Bicultural Education. U.S. Commission on Civil Rights Clearinghouse Publication 51, May, 1975.

U.S. General Accounting Office. Bilingual Education: An Unmet Need. Washington, DC: USGPO, 1976.

Zappert, L.T. and B.R. Cruz. Bilingual Education: An Appraisal of Empirical Research. Berkeley, CA: Bay Area Bilingual Education League/Lau Center, 1977.

Zirkel, P.A. "Bilingual Education Programs at the Elementary School Level: Their Identification and Evaluation." Bilingual Review/Revista Bilingüe 2(1-2):13-21, 1975.